UNIVERZITA KARLOVA V PRAZE
FILOZOFICKÁ FAKULTA
Ústav českého jazyka a teorie komunikace

Bakalářská práce

Jiří Svák

Možnosti chybové anotace češtiny nerodilých mluvčích

Possibilities of Error Annotation of Non-Native Speakers' Czech

Praha 2013

Vedoucí práce: prof. PhDr. Karel Šebesta, CSc.

Poděkování

Děkuji svému vedoucímu panu prof. Karlu Šebestovi za trpělivé vedení mé bakalářské práce, cenné připomínky a rady.

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

V Praze dne Jiří Svák

Klíčová slova

korpus, žákovský korpus, chybová anotace, distanční anotace, vícerovinná anotace, CZESL, FALKO

Key words

corpora, leaner corpora, error annotation, stand-off markup , multi-level annotation, CZESL, FALKO

Abstrakt

Bakalářská práce „Možnosti chybové anotace češtiny nerodilých mluvčích“ srovnává anotační systémy vybraných žákovských korpusů z pohledu chybové anotace. Pro srovnání byly zvoleny dva žákovské korpusy – český CZESL a německý FALKO. Oba korpusy používají distanční vícerovinný anotační model.

Práce je rozdělena na dvě části: teoretickou a praktickou.

V teoretické části jsou podrobně popsány oba vybrané korpusy a jejich anotační modely. Praktická část zpracovává anotaci žákovského textu v prostředí anotačních modelů obou korpusů.

Cílem práce je zdůraznit možná pozitiva i negativa vybraných anotačních formátů.

Abstract

Bachelor thesis „Possibilities of Error Annotation of Non-Native Speakers' Czech“ compares annotation systems of selected learners corpora from the perspective of error annotation. For the comparison, two learner corpora were chosen – Czech CZESL and German FALKO. Both corporas use stand-off multi-level annotation model.

The paper is divided into two parts: theoretical and practical.

In the theoretical part there is an in-depth description of both of selected corporas and their annotation models. The practical part presents annotation of pupil text processed in annotation models of both corporas.

The aim of this paper is to highlight possible strengths and weaknesses of selected annotation formats.

Obsah

1 Úvod

Bakalářská práce „Možnosti chybové anotace češtiny nerodilých mluvčích“ má za cíl porovnat anotační systémy vybraných žákovských korpusů s ohledem na jejich možnosti chybové anotace.

Práce je rozdělena na dvě hlavní části – deskriptivní (kap. 2., 3., 4. a 5.) a exemplifikační (kap. 6.).

Vzhledem k tomu, že chybová anotace korpusů má svůj základ v teorii chybové analýzy, bude druhá kapitola věnována stručnému uvedení do tohoto tématu. Následné kapitoly 3, 4 a 5 popisují vybrané korpusy a jejich anotační systémy a systémy metadat. Na základě těchto informací budou v kapitole 6 vybrané systémy srovnány a bude provedena exemplifikace. Ta bude probíhat jako anotace žákovského textu v každém z vybraných anotačních systémů, doplněná vysvětlujícím komentářem. Exemplifikace má za úkol vyznačit problémové případy v anotaci vybraných korpusů, a zdůraznit tak pozitivní i negativní stránky jednotlivých anotačních formátů.

2 Chybová analýza a chybová anotace

V 50. a 60. letech minulého století byla hlavní metodou pro popis a analýzu osvojování cizího jazyka kontrastivní analýza (James 1998, s. 4). Ta má své základy v behavioristické teorii nabývání jazyka, která tvrdí, že učení druhému jazyka je proces osvojování si určitých vzorců jazykového chování, které žák odezírá od rodilých mluvčích (Štindlová 2011, s. 8). Tato teorie upřednostňuje důležitost vstupu a upozaďuje vlastní mentální procesy žáka. Důsledkem behavioristické teorie pak byl rozvoj kontrastivní analýzy, která žákovské chyby porovnávala prismatem prvního jazyka, protože měla za to, že při učení druhému jazyku dochází k interferencím s jazykem rodným (Štindlová 2011, s. 9). Chyba byla brána jako defekt, který je nutný odstranit, aby nedocházelo k jeho nežádoucímu upevňování.

Kritikou behavioristické teorie a kontrastivní analýzy byl stvořen koncept mezijazyka. Žák není jen pouhým příjemcem vstupu, ale zastává roli aktivního činitele v učebním procesu – zkouší různé jazykové formy, chybuje či se dopouští omylů a následně se opravuje, čímž se přibližuje k jazykové normě dospělého mluvčího (Štindlová 2011, s. 22). Tato jazyková kompetence v průběhu procesu učení může být předmětem zkoumání stejně jako přirozený jazyk (James 1998, s. 6).

Ruku v ruce s teorií mezijazyka se vyvinula chybová analýza. Ta nebere chybu jako defektní záležitost, ale jako klíč, vodítko k porozumění jazykové akvizice (Štindlová 2011, s. 21). Zároveň se její těžiště přesouvá ze zkoumání možných interferenčních chyb prvního jazyka (jako u kontrastivní analýzy) na analýzu mezijazyka versus cílového jazyka (James 1998, s. 5).

2.1 Chybová analýza

Chybová analýza má tedy za úkol zkoumat chyby žákovského jazyka v souvislosti s jazykem cílovým.

Základy chybové analýzy popsal S. P. Corder ve své práci *The significance of learners' errors*: [1]

[1] Zpracováno podle (James 1998, s. 12–13)

- Učení se druhému jazyku má své paralely v osvojování prvního (mateřského) jazyka.
- Při učení druhého jazyka si žák vytváří svůj vlastní jazykový systém („in-built syllabus“). Chyby jsou pak odchylky od tohoto systému, podobně jako je tomu v případě mateřského jazyka.
- Chyby by se měly rozdělovat na „errors“ (chyby systémové) a „mistakes“ (nesystémové chyby).[2]
- Zkoumání chyb je důležité z tří důvodů:

  chyby ukazují učiteli, co je třeba učit;

  chyby ukazují učiteli, jak učení jazyka probíhá;

  chyba je projevem procesu testování žákových hypotéz ohledně cílového jazyka.

Proces chybové analýzy je možné rozdělit do několika fází (podle James 1998, s. 90 a dále): sběr dat, identifikace chyby, vymezení chyby, popis chyby, klasifikace chyby a explanace chyby.

Identifikace chyby se děje na základě srovnání nestardardní formy s očekávanou formou jazyka dospělého rodilého mluvčího.

Při dalším kroku, popisu a klasifikaci chyby, se předpokládá vytvoření systému kategorizace chyb neboli chybové taxonomie. K tomu je možné přistupovat z několika hledisek. Buďto je možné založit chybovou taxonomii na popisu změn v povrchové realizaci textu (vynechání, přidání, špatné užití, změna slovosledu, spojení), anebo na základě lingvistických kategorií, či oba tyto přístupy kombinovat (James 1998, s. 102–114).

Fáze explanace je založena na tzv. komparační taxonomii, která se skládá ze čtyř tradičních kategorií: interlingvální chyby na základě mezijazykového transferu, intralingvální chyby na základě cílového jazyka, učební neboli tzv. vynucené chyby a chyby vyplývající z uplatňovaných komunikačních strategií (Štindlová 2011, s. 29–30).

[2] Podle (James 1998, s. 76–79): Nesystémová chyba (mistake) je taková chyba, jejíž existenci si mluvčí dokáže uvědomit a posléze ji i opravit. Pokud je však chyba důsledkem vlastních pravidel žákovského mezijazyka, jedná se o chybu systémovou (error).

2.2 Kritika chybové analýzy[3]

Tím, že se chybová analýza zaměřuje přednostně na analýzu chyb, jedním z hlavních nedostatků byla podle jejích kritiků absence uceleného obrazu žákovského jazyka.

Další výtky byly směřovány k nemožnosti chybové analýzy vypořádat se strategií vyhýbání[4] a rovněž byla kritizována samotná metodologie deskripce chyb pro její neobjektivnost a neuspořádanost. Vypracování spolehlivější typologie chyb by však vyžadovalo rozsáhlou databanku jazykových dat (jinými slovy korpus), kterých se v té době nedostávalo.

2.3 Chybová analýza s oporou korpusu[5]

S rozvojem korpusové lingvistiky bylo možné využít její vlastnosti pro potřeby zkoumání akvizice cizích jazyků. Pro chybovou analýzu to znamenalo možnost překonat výše uvedené nedostatky.

První výhodou je, že s možností elektronizace a počítačového zpracování textů lze snížit problém nedostatku kvalitních jazykových dat. Soubory dat mohou dosahovat značných velikostí a stále zůstat spravovatelné, což v době před nástupem počítačů nebylo možné. Data mohou být taktéž snadno systematicky strukturovány a opatřovány doplňujícími informacemi, což zvyšuje jejich spolehlivost a rozšiřuje možnosti jejich využití. A v neposlední řadě mohou obsahovat žákovské chyby v kontextu. Uvedené výhody negují připomínky k nemožnosti zachytit komplexní žákovský jazyk.

Další výtka, totiž nespolehlivost chybové typologie, je za pomoci korpusů rovněž řešitelná. Výzkumníci jsou schopni přeformulovat chybové taxonomie na základě objektivnějšího korpusového výstupu. Podle autorů článku *Error tagging systems for learner corpora* (Díaz-Negrillo a Fernández-Domínguez 2006, s. 86) však stále zůstává problém standardizace těchto chybových taxonomií.

---

[3] Podle (Štindlová 2011, s. 33–34).

[4] Strategie nerodilého mluvčího, který se při projevu v cizím jazyce záměrně vyhýbám jevům, které jsou pro něj obtížné, nebo které neovládá.

[5] Zpracováno podle (Díaz-Negrillo a Fernández-Domínguez 2006, s. 84–86).

3 Vybrané korpusy a materiál

3.1 Předmět srovnání

Pro potřeby srovnání byly vybrány dva korpusy: korpus češtiny jako cizího jazyka, neboli Corpus of Czech as a second language, zkráceně CZESL, a chybový korpus němčiny jako cizího jazyka, neboli Fehlerannotiertes Lernerkorpus, zkráceně FALKO. Oba korpusy jsou koncipovány jako multilingvální z pohledu výchozího jazyka a monolingvální z pohledu jazyka cílového – v prvním případě je cílovým jazykem čeština, v druhém případě němčina.

Korpus FALKO byl vybrán, protože mezi žákovskými korpusy němčiny jako cílového jazyka užívá vícerovinný distanční anotační model. V rámci žákovských korpusů byl FALKO prvním korpusem, který tento typ anotace zavedl.

Korpus FALKO byl inspirací pro korpus CZESL, který zde zastupuje jediný chybově anotovaný korpus češtiny jako cizího jazyka. Podrobné popisy obou korpusů, které jsou zároveň východiskem pro samotné srovnání, jsou zpracovány v kapitolách 4 a 5 této práce.

3.2 Charakteristika materiálu

Srovnání možností anotace obou anotačních modelů bude probíhat na vybraném textu nerodilého mluvčího. Text bude anotován anotačním systémem CZESLu a paralelně s ním také anotačním systémem subkorpusů FalkoEssay a FalkoSummary.

Tímto způsobem bude možné konfrontovat jednotlivé anotační modely na základě reálných dat, ne pouze na základě teoretického porovnání.

Jedná se o text, který byl sesbírán v rámci tvorby korpusu CZESL a nyní je jeho součástí. Text byl vybrán s ohledem na přiměřenou chybovost – korpus FALKO pracuje jen s žákovskými texty vyšších úrovní, některé začátečnické chyby textů korpusu CZESL by mohly být v prostředí Falka neanotovatelné.

Text je k dispozici pod identifikátorem HRD_XY_026_t_1[6] skrze webové rozhraní

[6] *CQL* dotaz <doc name="HRD_XY_026_t_1" />

NoSketch Engine korpusu CZESL-PLAIN na stránkách ÚČNK[7].

Část, která bude anotována, citujeme zde:

Můj nejhorší den v životě

Studovala jsem v střední škole, měla jsem nejhorší den. Ráno jsem stavala pozdě, protože můj budík nefungoval. Stihnula jsem do školy autobusem zase měla jsem dopravní zacpu, čekala jsem kolem 1 hod. Když jsem jela do školy měli jsme výučovaní už končil. byla jsem špatná. Ale jsem neměla dobrý napad, a jenom jsem cvíčila hudební stroje, protože jsem měla vyučov na hudební stroje.

[7] Viz http://korpus.cz/hledat_v_cnk.php

4 Popis korpusu CZESL

Žákovský korpus češtiny[8] (neboli CZESL – *Corpus of Czech as a Second Language*) vznikl jako jeden z výstupů projektu Inovace vzdělávání v oboru čeština jako druhý jazyk. Řešitelem je Technická univerzita v Liberci ve spolupráci s Karlovou univerzitou, Asociací učitelů češtiny jako cizího jazyka a dalšími pracovišti. Korpus je zároveň zahrnut do širšího projektu akvizičních korpusů AKCES.[9]

CZESL je koncipován jako korpus psané i mluvené češtiny jako cílového jazyka (Šebesta a Škodová 2012, s. 29). Z pohledu prvního jazyka je to korpus multilingvální (Šebesta 2010, s. 27). Je zacílen na mluvčí na území Čech, kteří užívají češtinu jako cizí jazyk. Vzhledem ke svému pedagogickému účelu se zaměřuje primárně (ale nikoli výlučně) na jazykové a kulturní menšiny z teoretického i praktického hlediska nejrelevantnější – rusky a jinými slovanskými jazyky hovořící cizince, vietnamskou komunitu a romskou komunitu (Šebesta 2010, s. 27; k romské komunitě viz Šebesta a Škodová 2012, s. 109). Neopomíjí však i méně početné skupiny cizinců.

Teoreticky by se tedy dal textový materiál rozdělit na tři skupiny: jazyky češtině blízké (ruština, ukrajinština, polština), neindoevropské jazyky (především vietnamština, ale i arabština) a zbývající jazyky indoevropského původu (například francouzština, němčina apod.) (Štindlová 2011, s. 99–100; Šebesta 2010, s. 27). Zvláštním případem jsou pak romští mluvčí češtiny ze sociálně vyloučených lokalit – jejich texty tvoří samostatný subkorpus ROMI, který má některé specifické atributy (především co se týče metadat; o tom kapitola 4.2) (ANON. 2013c).

V současné době je korpus ve své neanotované podobě (jako CZESL-PLAIN) uveřejněn na stránkách Ústavu Českého národního korpusu při Filozofické fakultě Univerzity Karlovy.[10] Velikost korpusu ukazuje názorně tabulka 1:

[8] Viz http://utkl.ff.cuni.cz/learncorp (ANON. 2013a) a http://www.c2j.cz/projekt-a-realizacni-tym (ANON. 2013h).

[9] Viz http://akces.ff.cuni.cz/ (ANON. 2013b).

[10] Viz http://korpus.cz/struktura.php (ANON. 2013d).

Tabulka 1: Velikost korpusu CZESL[11]

Typ textů	Počet textů	Počet pozic (slova +interpunkce)
ciz – eseje cizinců	8 863	1 314 901
kval – odborné kvalifikační práce	176	731 816
rom – slohové práce romských žáků	4 420	428 161
CELKEM	13 459	2 474 878

Uvedené množství v tabulce 1 je vyjádřeno v tokenech[12], což jsou všechny jednotlivé výskyty slovního tvaru, čísla či interpunkčního znaménka (jak vysvětluje i záhlaví tabulky). Udávaná velikost v počtu slov (tedy bez interpunkce) je pak asi 2 miliony (Šebesta a Škodová 2012, s. 29).

Anotovaná část korpusu je přístupná prostřednictvím rozhraní *SeLaQ* (jež je momentálně ve vývoji) skrze odkaz na webových stránkách Ústavu teoretické a komputační lingvistiky Filozofické fakulty Univerzity Karlovy.[13]

Mluvená složka prozatím nebyla uveřejněna a je postupně zpracovávána.

4.1 Jazyková data korpusu

Jazykový materiál korpusu můžeme pro potřeby popisu rozdělit na písemný a mluvený. Písemný materiál představují texty z výuky češtiny jako cizího jazyka (především žákovské eseje, v menší míře i kvalifikační práce) (Šebesta a Škodová 2012, s. 29). Mluvený materiál je záznamem elicitovaných interview na obecná témata (Šebesta 2010, s. 28). Texty jsou obsahově různorodé, pro zařazení do korpusu nejsou zavedena žádná tematická či žánrová omezení (Štindlová et al. 2011, s. 211).

Korpus se zaměřuje na všechny úrovně mluvčích, obsahuje tedy data jak od pokročilých studentů, tak i od začátečníků (čímž se odlišuje od většiny světových žákovských korpusů – ty začátečnické úrovně cílového jazyka nezpracovávají, především kvůli jejich anotační a emendační náročnosti) (Štindlová 2011, s. 100). Texty kvalifikačních prací (bakalářské, diplomové i doktorské) tvoří v konečné fázi kvůli své specifičnosti subkorpus, viz tabulka 1 (Štindlová et al. 2011).

[11] Převzato z http://www.korpus.cz/czesl-plain.php (ANON. 2013c)

[12] Předkládaná definice *tokenu* – viz nápovědu u údaje o velikosti korpusu CZESL-PLAIN ve vyhledávacím rozhraní NoSketch Engine (ANON. 2013f).

[13] Viz http://utkl.ff.cuni.cz/learncorp (ANON. 2013a).

4.2 Metadata

Korpus CZESL se snaží o co nejširší využití metadat. Informace o osobě mluvčího, o situaci a podmínkách sběru a o textu samotném se přiřazují ve formě parametrů. Jednotlivé parametry pak mohou nabývat více hodnot. Přehled parametrů znázorňuje tabulka 2.

Tabulka 2: Zaznamenávané parametry korpusu CZESL[14]

Parametry spojené s textem (4)		
Médium		
Převažující slohový postup		
Téma		
Typ tématu		
Parametry spojené se situací vzniku a sběru (16)		
Obecně/psané projevy	Pouze u mluvených projevů	Pouze u sociokulturně znevýhodněných komunit a srovnávací skupiny rodilých mluvčích
Je zadán slohový postup?	Forma projevu	Místo sběru
Je zadáno téma?	Počet mluvčích	
Je zadán rozsah?	Formálnost situace	
Je zadán časový limit?	Míra připravenosti	
Přípravná aktivita (zvl. u psaných projevů)	Sběrač – první jazyk	
Je projev součástí zkoušky?	Sběrač – věk	
Je povoleno užití slovníku nebo jiné pomůcky (u psaných projevů)?	Sběrač – vztah k žákovi	
Prostředí pořízení materiálu		

[14] Podle (Šebesta a Škodová 2012, s. 30–32).

Parametry spojené s žákem (24)		
Obecně	Pouze u nerodilých mluvčích	Pouze u sociokulturně znevýhodněných komunit a srovnávací skupiny rodilých mluvčích
Pohlaví	Skupina prvních jazyků	Navštěvovaná škola
Věk	Studium češtiny – místo	Třída/ročník
První jazyk	Studium češtiny – doba	Mluví žák romsky?
Znalost dalších jazyků	Studium češtiny – intenzita	Mluví někdo blízký romsky?
	Studium češtiny – učebnice	První jazyk
	Znalost češtiny podle SERR	Doma mluví
	Bilingvnost?	Bydliště – velikost sídla
	Délka pobytu v ČR	Bydliště – region
	Umí někdo v rodině česky?	Bydliště – nářeční oblast
		Bydliště – sociálně vyloučená lokalita?
		Poznámky sběrače/učitele

V tabulce 2 vidíme diferenciaci parametrů podle typu projevu a skupiny mluvčích. K textům sociokulturně znevýhodněných komunit (tvoří subkorpus ROMI) a srovnávací skupiny rodilých mluvčích jsou přiřazovány odlišné parametry, vzhledem k tomu, že se nejedná o mluvčí, kteří užívají češtinu jako cizí jazyk, anebo jsou parametry modifikovány. U subkorpusu ROMI je také kladen důraz na parametry, které mapují vliv sociálního prostředí (Šebesta a Škodová 2012, s. 30).

V současné době nemá korpus CZESL-PLAIN metadata doplněna, na jejich kompletaci se však pracuje.

4.3 Anotace

Korpus CZESL je anotován jak morfologicky, tak i chybově. Nejprve probíhá manuální chybová anotace, poté je provedena automatická anotace, při které se přiřadí lingvistické specifikace k jednotlivým slovům a doplní chybové značky, u nichž není potřeba asistence anotátora (Štindlová 2011, s. 120).

Anotace proběhla zatím jen na malé části korpusu. Ačkoli je zpracována koncepce anotace, jsou vytvořeny potřebné nástroje a připraven anotační tým, jde o práci velmi náročnou, tudíž postupuje velmi pomalu.

4.3.1 Chybová anotace

Při tvorbě korpusu CZESL byly na anotační formát kladeny tři základní požadavky, jak je formulovala B. Štindlová v monografii *Čeština – cílový jazyk a korpusy* (Šebesta a Škodová 2012, s. 62):

„1. [anotační] schéma musí být zvladatelné pro anotátory,

2. taxonomie nemůže být příliš rozsáhlá, ale zároveň musí být dostatečně informativní, tj. musí umožňovat dostatečně podrobné zachycení chyb,

3. taxonomie by měla umožňovat budoucí rozšiřování.“

Zároveň bylo nutné vzít v potaz, že čeština je vysoce flektivní jazyk s volným slovosledem. Jednorovinným anotačním formátem (viz kapitola 5.3, s. 34) se obtížně zachycují opravy slovosledu či opravy nesousedících nebo nespojitých řetězců (Štindlová 2011, s. 118).

Anotovaný korpus by taktéž měl poskytovat dostatečně relevantní a podrobné informace k pozdějšímu statistickému/pedagogickému využití, proto je vhodné zachovat co nejvíce informací, které přispívají k identifikaci a explikaci chyby. Z tohoto důvodu preferuje korpus CZESL postupnou anotaci na více rovinách, kdy je možné zaznamenat průběh emendace, a zachovat tak údaje, které k opravě vedly, a taktéž sledovat vztahy mezi jednotlivými tokeny (Šebesta a Škodová 2012, s. 63–34; Štindlová 2011, s. 118).

Pro korpus CZESL byl tedy vytvořen specifický vícerovinný anotační formát, který byl inspirován korpusem FALKO (viz kapitola 5 této práce).

Anotace korpusu FALKO je založena na vícerovinné distanční architektuře s pohyblivým počtem rovin, přičemž lze formulovat více cílových hypotéz (Štindlová 2011, s. 118). CZESL naproti tomu zvolil systém, který počítá pouze s jedinou cílovou hypotézou. Taktéž počet rovin není pohyblivý – byly zvoleny roviny tři (R0, R1, R2).

Rovina R0 obsahuje původní přepsaný text. Na rovině R1 se v textu opravují chyby na úrovni jednotlivých slovních tvarů bez ohledu na kontext, na rovině R2 se pak opravují chyby kontextově motivované (valence, slovosled, determinace apod.) (Šebesta a Škodová 2012, s. 64).

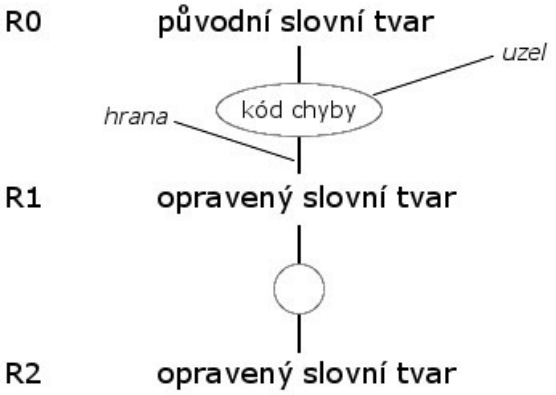
Toto rozdělení tedy reflektuje i lingvistická kritéria – na rovině R1 se opravují především morfologické a ortografické chyby, chyby na úrovni vztahů mezi větnými

jednotkami se opravují na rovině R2.

Rozvržení umožňuje postupnou anotaci a explicitní vyjádření vztahů mezi slovními tvary na jednotlivých rovinách (Štindlová 2011, s. 118), přičemž náročnost na anotátora zůstává únosná.

Základní anotační schéma znázorňuje následující obrázek:

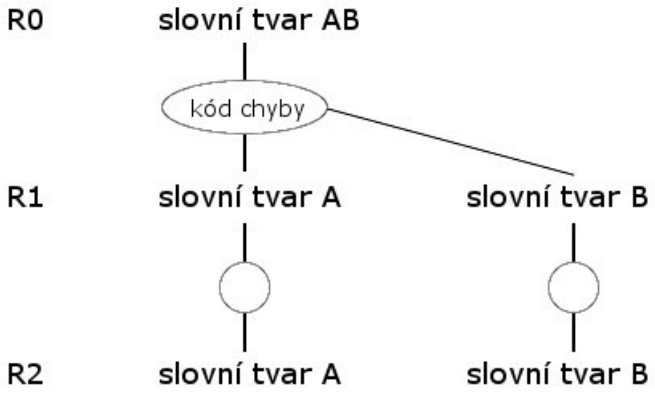
Obrázek 1: Anotační schéma korpusu CZESL[15]



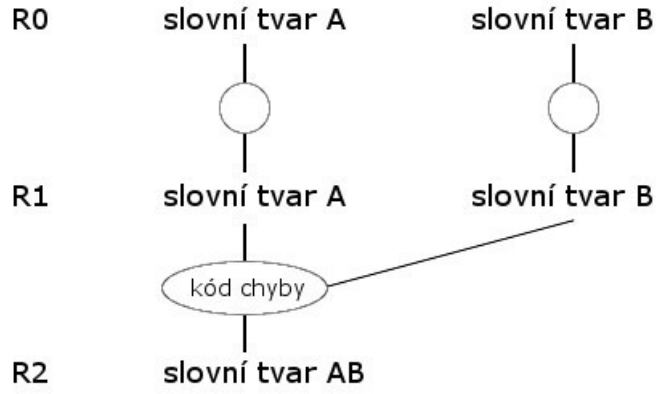
Na obrázku 1 vidíme zmiňované rozložení na 3 roviny. Každý slovní tvar (token) je spojen hranou s odpovídajícím slovním tvarem (tokenem) na nižší rovině, v uzlu této hrany je pak možno přiřazovat kód chyby (Rosen a Štindlová 2012, s. 7). Obrázek 1 je příkladem uspořádání, kdy každému ze slovních tvarů odpovídá na nižší rovině právě jeden slovní tvar. Anotační formát korpusu CZESL však umožňuje přiřadit na další rovině k jednomu slovnímu tvaru i více tokenů, anebo více tokenů spojit v jeden, či naznačit jiné složitější vztahy (viz obrázek 2, 3 a 4). Takovémuto uspořádání se říká pavouk (Rosen a Štindlová 2012, s. 9). Pavouk umožňuje zachovávat korespondence napříč rovinami při rozdělování a spojování slovních tvarů.

[15] Vytvořeno v programu GIMP (Simončič 2013).

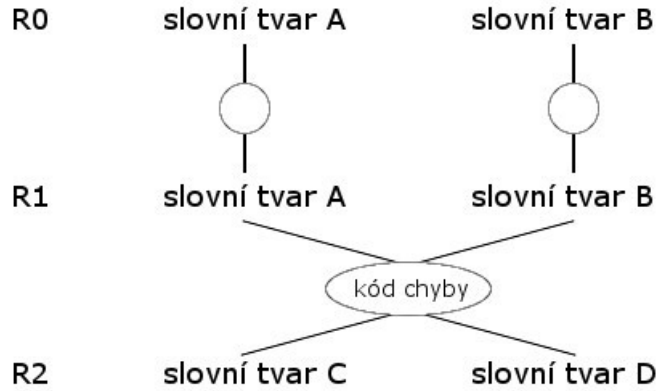
Obrázek 2: Rozdělení slovního tvaru na rovině R1[16]



Obrázek 3: Spojení slovního tvaru na rovině R2[17]



Obrázek 4: Záměna dvousloví za jiné dvousloví[18]



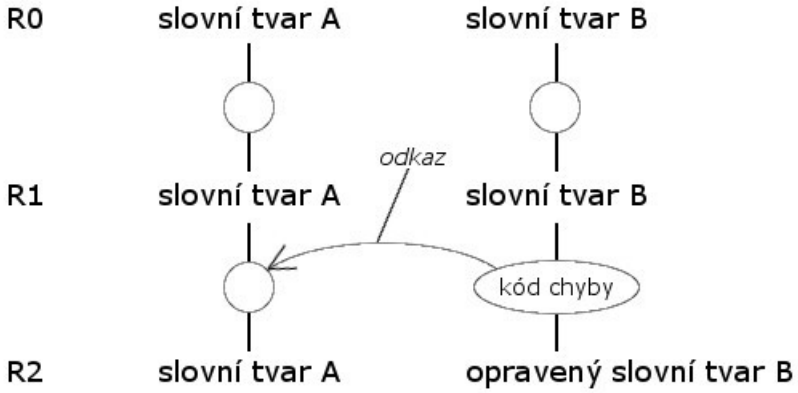
[16] Vytvořeno v programu GIMP (Simončič 2013).

[17] Vytvořeno v programu GIMP (Simončič 2013).

[18] Vytvořeno v programu GIMP (Simončič 2013).

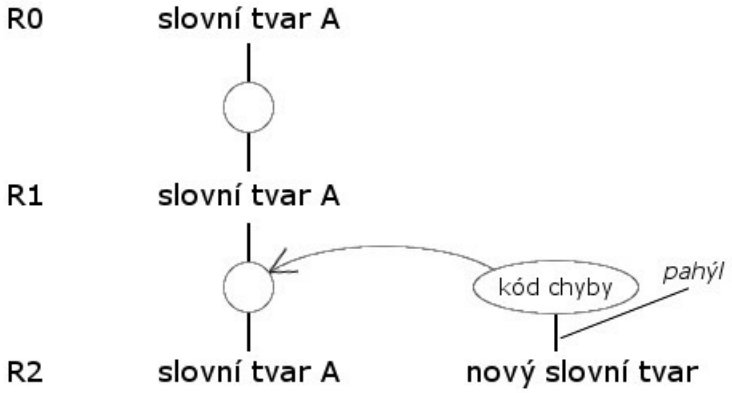
Rovněž je možné vést spojnice mezi uzly jednotlivých slovních tvarů na téže rovině. Spojnice mají odkazující charakter, vedou většinou od opraveného slovního tvaru ke slovu motivujícímu k opravě chyby (viz obrázek 5). Podle tohoto způsobu užití se nazývají odkazy (Rosen a Štindlová 2012, s. 10). Odkazy se vkládají výlučně na rovině R2, kde mají své opodstatnění v naznačování vztahů u kontextově motivovaných oprav.

Obrázek 5: Odkazování na rovině R2[19]



Kromě výše popsaných poměrů mezi tokeny na jednotlivých rovinách (1:1 a 1:2, respektive 2:1) existuje i stav, kdy na další rovině neodpovídá slovnímu tvaru žádný token. Takovéto konformaci se říká pahýl a vzniká při vymazání celého tokenu či při jeho přidání (Rosen a Štindlová 2012, s. 9–10). Na pahýl lze přiřadit kód chyby a vést od něj (či k němu) odkaz (viz obrázek 6).

Obrázek 6: Pahýl[20]



[19] Vytvořeno v programu GIMP (Simončič 2013).

[20] Vytvořeno v programu GIMP (Simončič 2013).

4.3.2 Taxonomie chybové anotace

Chybová taxonomie byla vytvořena se zřetelem na plánovaný cíl korpusu: poskytnout pedagogickým pracovníkům platformu pro výzkum akvizice jazyka nerodilých mluvčích češtiny. Jedná se o účel velmi široký, proto byla zvolena taxonomie snažící se o komplexní kategorizaci chyb (Štindlová 2011, s. 80).

Značky (tagy, chybové kódy) jsou definovány na základě lingvistických kategorií (například *agr* – chyba ve shodě, *rflx* – chyba v reflexivním výrazu) a povrchové realizace (formální kritéria; například *miss* – chybějící slovo, *wo* – chybný slovosled) (Šebesta a Škodová 2012, s. 65).

Užití značek podmiňuje anotační rovina. Značky popisující chyby izolovaných slovních tvarů se přiřazují na R1, na R2 se přiřazují značky pro chyby kontextově motivované. Tagy vyhovující oběma kritériím se mohou přiřadit na rovinu R1 i R2.

Značky se aplikují na uzel hrany spojující slovní tvar mezi dvěma rovinami (viz anotační schémata v kapitole 4.3.1). Mají formu textového identifikátoru, tzv. domény, která je u některých značek ještě rozšířena pro bližší specifikaci o další identifikátor, tzv. typ (Štindlová 2011, s. 119). Pro ilustraci uvádíme následující příklad:

incorInfl

doménaTyp

Příklad zobrazuje kód, který označuje nesprávný tvar (doména *incor*) a blíže ho specifikuje jako chybu ve flexi (typ *Infl*).

Na jeden uzel je možné umístit i více chybových značek, pokud slovní tvar obsahuje více typů chyb.

Korpus CZESL má pro manuální chybovou anotaci k dispozici 22 tagů (Rosen a Štindlová 2012, s. 60). Podrobný seznam chybových kódů s doplňujícími informacemi je k nalezení v příloze (viz příloha 1, s. 65 této práce).

Na rovině R1 se manuálně značkují chyby týkající se nesprávného tvaru ve slovním základu a ve flexi, dále chyby spočívající ve špatném rozdělení či spojení slov.[21] Označují se autorské novotvary, cizojazyčná či jinak obtížně identifikovatelná a neznámá slova, dále pak výplňková slova a tvary slangové, nářeční, knižní a obecně české. Poslední čtyři

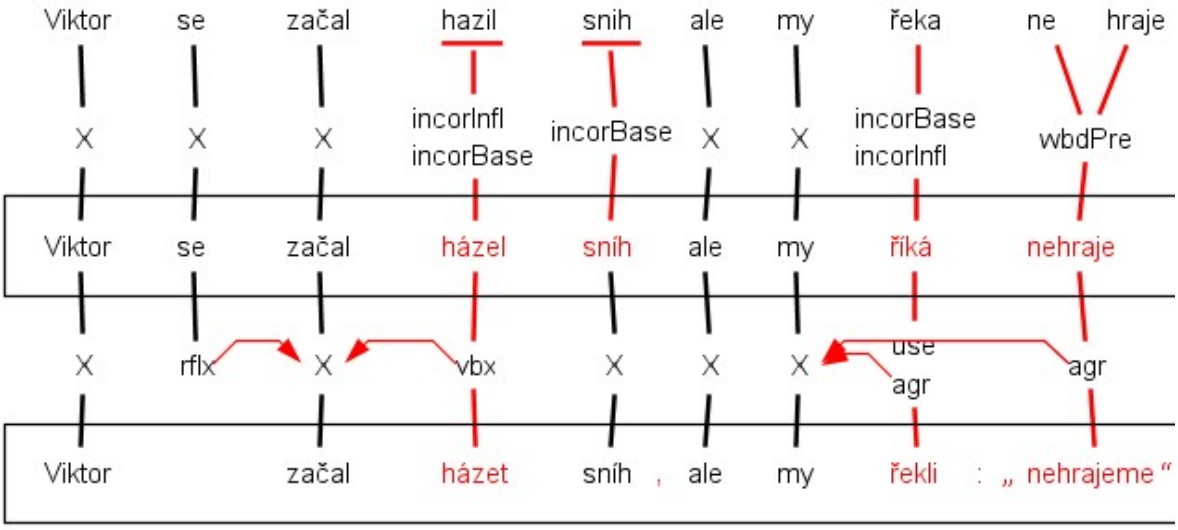
[21] Tento a následující odstavec jsou zpracovány podle (Štindlová 2011, s. 118–120).

jmenované značky je možno použít i na rovině R2. Atypický je i tag *problem*, který označuje jakýkoli problémový element a je rovněž možné ho použít na obou rovinách. Zároveň se jedná o tag doplňkový, tedy nepřiřazuje se sám o sobě, ale pouze s jinou chybovou značkou. Doplňkový charakter má i kód *flex*, který se přiřazuje na rovině R1 k novotvarům a cizojazyčným slovům, pokud je u nich přítomna česká flexe.

Na rovině R2 se manuálně značkují chyby ve shodě, syntaktické závislosti, zájmenném odkazu, chyby týkající se reflexních výrazů, tvarů složeného predikátu a analytického slovesa obecně. Dále zde patří chyby v negaci a ve špatném užití lexika či gramatické kategorie. Na rovině R2 se taktéž určují chyby v povrchové realizaci (nadbývající anebo chybějící slovo, špatný slovosled). Rovina R2 rovněž zavádí dvě doplňkové značky – *disr* a *sec*. První označuje rozvrácenou konstrukci, druhá se používá v případě, že oprava nějakého slovního tvaru byla vynucena jinou dřívější opravou. Řečeno příkladem: pokud je z nějakého důvodu změněn pád řídícího substantiva v jinak správně utvořené přívlastkové konstrukci, změní se tím i pád přívlastku, aby zůstala zachována shoda. Přívlastek pak bude označen tagem *sec*, protože bez změny řídícího substantiva by k emendaci nedošlo („uvařil dobrá polévka“ > „ uvařil dobrou *(agr+sec)* polévku *(dep)*“).

Informace z podkapitol 4.3.1 a 4.3.2 ilustrujeme souborně na reálných datech – následující obrázek 7 zobrazuje finální podobu anotace a emendace části jednoho souvětí. Jedná se o upravený výřez z programu *feat* (program pro manuální anotaci a její správu, viz kapitolu 4.4).

Obrázek 7: Ukázka anotace[22]



Při chybové anotaci platí tzv. zásada minimálního zásahu anotátora (Rosen a Štindlová 2012, s. 12). Anotátor by měl přihlížet k pravděpodobné intenci autora, nicméně by se měl snažit text co nejméně upravovat. Výsledkem by měl být „[…] souvislý a gramaticky správný text, ale bez nároků na stylistickou vytříbenost.“ (Šebesta a Škodová 2012, s. 65)

4.3.3 Automatická anotace

Probíhá dvojího druhu – nejprve automatická lingvistická anotace, poté automatické doplnění a rozšíření chybových značek.

4.3.3.1 Lingvistická anotace

Lingvistická anotace přiřazuje ke slovům na obou rovinách lemma, slovní druh a morfologické kategorie (Štindlová et al. 2011, s. 221). Průběh procesu se však na jednotlivých rovinách liší.

Na rovině R2 se provede morfologická analýza, při které se u každého slova určí všechny možné kombinace lemmat a gramatických kategorií, kterých může nabývat (Jelínek 2008, s. 14). Poté je aplikována disambiguace, což je proces, který slovu přidělí ze všech údajů získaných v předchozím kroku pouze ty, které jsou s ohledem na kontext

[22] Věta je k nalezení v korpusu CZESL-PLAIN pod označením HRD_YO_090_t_1 (je možné ji zobrazit pomocí CQL dotazu „<doc name="HRD_YO_090_t_1" />“). Její znění bylo pro názornější ukázku možností anotace pozměněno.

správné (Jelínek 2008, s. 14). Výsledkem je, že každé slovo na rovině R2 je jednoznačně lingvisticky anotováno.

Na rovině R1 nevede lingvistická anotace k jednoznačnosti, jako je tomu na R2. Slovo na R1 je gramaticky správné pouze jako osamocený tvar bez ohledu na kontext, proto zde nelze provést disambiguaci. Postupuje se tedy následovně (podle Šebesta a Škodová 2012, s. 78):

1. slovo, které je stejné jako jeho protějšek na rovině R2, přebírá tagy slova z roviny R2
2. pokud se tvar slova na rovině R1 od R2 odlišuje, ale shoduje se v lemmatu, přiřadí se pouze značky relevantní pro oné lemma
3. pokud není stejné ani lemma, přiřadí se všechny možné morfologické tagy (provede se tedy pouze prvotní morfologická analýza)

Tagy určující morfologické kategorie slova mají stejný formát jako tagy korpusů ČNK – tzv. poziční systém (Šebesta a Škodová 2012, s. 78). Jedná se o řetězec až 16 pozic, přičemž každá z nich představuje určitou morfologickou kategorii (anebo její upřesňující podkategorii) (ANON. 2013g). Pokud slovo této kategorie nenabývá či se neurčuje, pozice zůstane prázdná. Úplný seznam pozic s jejich hodnotami je uveřejněn na stránkách ČNK: http://korpus.cz/bonito/znacky.php (ANON. 2013g). Pro zběžný náhled na poziční systém postačí obrázek programu pro generování morfologických tagů ČNK, který vytvořila Hana Skoumalová; viz obrázek 8.

Obrázek 8: Znázornění pozic v morfologickém tagu ČNK[23]

Zkratky: 1/POS – slovní druh; 2/SUBPOS – bližší určení slovního druhu; 3/GENDER – jmenný rod; 4/NUMBER – číslo, 5/CASE – pád; 6/POSSG – přivlastňovací rod; 7/POSSN – přivlastňovací číslo; 8/PERSON – osoba; 9/TENSE – čas; 10/GRADE – stupeň; 11/NEG – negace; 12/VOICE – aktivum/pasivum; 13/res. a 14/res. – nepoužito; 15/VAR – varianta, stylový příznak; 16/ASPECT - vid

POS	SUBPOS	GENDER	NUMBER	CASE	POSSG	POSSN	PERSON	TENSE	GRADE	NEG	VOICE	res.	res.	VAR	ASPECT
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
												-	-		

Jako příklad uveďme slovo „matčiných“. Jedná se o přivlastňovací adjektivum ženského rodu v genitivu. Poziční tag v korpusu by vypadal následovně:

AUFP2F--------. *

A = adjektivum, U = adjektivum přivlastňovací (-ův/-in), F = femininum, P = plurál, 2 = genitiv, F = femininum (Skoumalová 2013). Zbylé kategorie se neurčují.

4.3.3.2 Doplňková chybová anotace

Automatická chybová anotace doplňuje takové značky, u kterých není nutné zpracování anotátorem.

Na rovině R1 se jedná o případy, u nichž k správné anotaci postačuje pouze porovnání originálního tvaru na rovině R0 s tvarem emendovaným (Šebesta a Škodová 2012, s. 79). Tímto způsobem se dají efektivně anotovat chyby v povrchové realizaci: chyby v diakritice, v užití malých/velkých písmen, chyby ve spodobě znělosti, ortografické chyby, chyby v neproběhlé palatalizaci, v protetickém v/j, či obecněji chyby v nadbývajících/chybějících/zaměněných znacích anebo řetězcích znaků (Šebesta a Škodová 2012, s. 80–81). Pro podrobnou typologii chyb odkazujeme na strany 79–81

[23] Obrázek programu byl převzat z webové stránky http://utkl.ff.cuni.cz/~skoumal/morfo/ (Skoumalová 2013).

v monografii *Čeština – cílový jazyk a korpusy* (Šebesta a Škodová 2012).

Na rovině R2 se k chybám povrchové realizace přiřazují značky automaticky také. Jedná se případy špatného slovosledu (*wo*), vynechání slova (*miss*) či jeho nadužití (o*dd*).

Automatická anotace má také za cíl rozšířit anotaci prováděnou manuálně. Při manuální anotaci je u některých typů chyb vyžadováno pouze základní určení, doplnění pak už probíhá automaticky.

Na rovině R1 dochází k zpřesnění tagu *wbd* (chybná hranice slov), kdy se označí, jestli se jedná o chybu v rozdělení či spojení slov (Šebesta a Škodová 2012, s. 81).

Na rovině R2 se doplňují chybové značky porovnáváním tagů přidělených při lingvistické anotaci (Štindlová 2011, s. 120). Doplňuje se značka *rflx* u chyb ve shodě (*agr*), syntaktické závislosti (*dep*) a zájmenném odkazování (*ref*), pokud se oprava týká reflexiva, a zpřesňuje se také určení chyby v analytickém slovesném tvaru a složeném přísudku (*vbx*) (Šebesta a Škodová 2012, s. 82).

V závěru se také identifikují a označují frazémy, idiomy a ustálené kolokace (Šebesta a Škodová 2012, s. 83–84)

Pro podrobný přehled znova odkazujeme na monografii *Čeština – cílový jazyk a korpusy* (Šebesta a Škodová 2012, s. 81–84).

4.4 Průběh anotace

Proces anotace má několik fází. Nejprve se shromážděné texty přepíší v textovém editoru do elektronické podoby, zpracují se do formátu html, kde se označí základní meta-informace (autorovy poznámky, přepisy apod.)[24], a zkonvertují do formátu vhodného pro zpracování v programu *feat*[25], neboli Flexible Error Annotation Tool, viz obrázek 9. (Šebesta a Škodová 2012, s. 70). Takto zpracovaný text je nahrán do systému *Speed*[26], což je program pro komplexní správu souborů anotačního procesu. Umožňuje odesílat texty určené k anotaci a hotové texty zase přijímat a předávat dále supervizorům ke kontrole a opravě (Šebesta a Škodová 2012, s. 72). Veškerá interakce probíhá v rámci anotačního

[24] Přepisovači se řídí manuály *Manuál pro přepis psaných materiálů* (Hrdličková 2011) a *Doplnění k manuálu 3* (Rosen 2011).

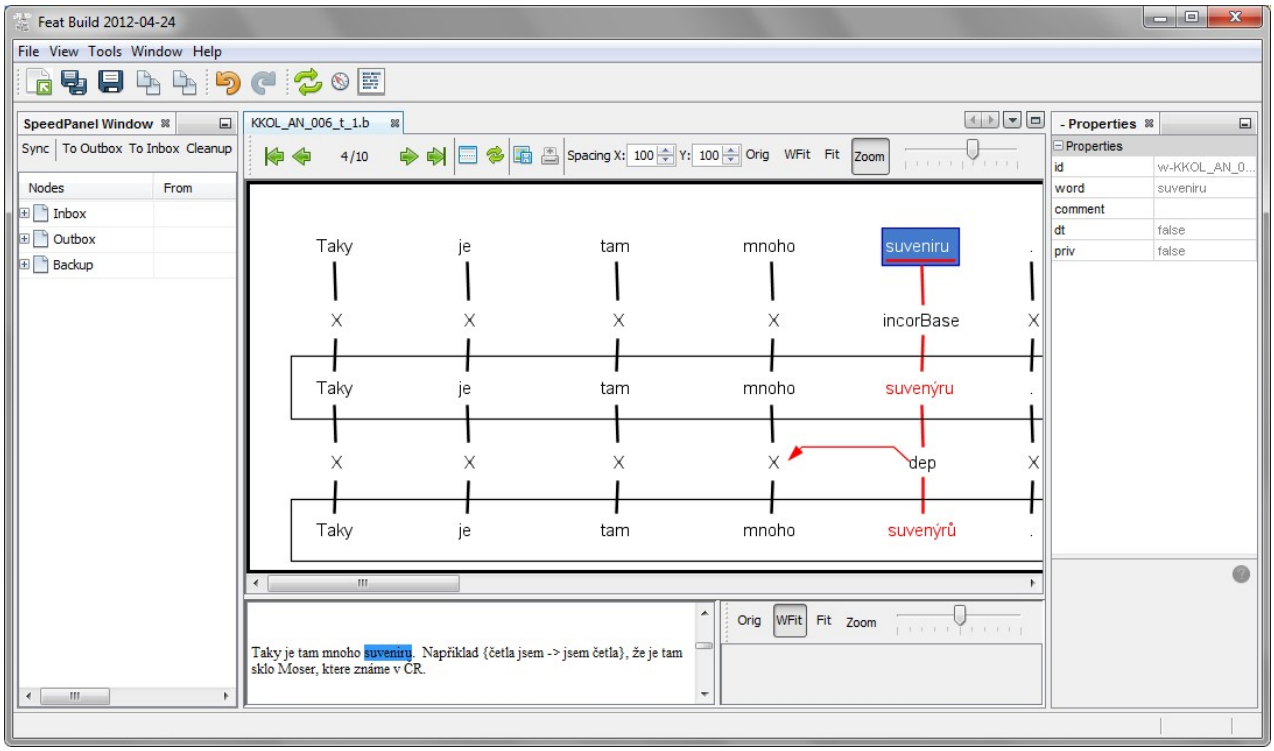
[25] Program *feat* je dostupný na adrese http://ufal.mff.cuni.cz/~hana/feat.html, viz (Hana 2012).

[26] Webové stránky korpusového manažeru *Speed* jsou k nalezení na adrese http://speed.aspone.cz/.

programu *feat*, protože systém *Speed* je do něj integrován.

Anotace je vždy částečně interpretací anotátora, proto se při zpracování textů musí počítat s tím, že se jednotlivé anotace budou lišit (Štindlová 2011, s. 121) [27]. Aby se zamezilo alespoň zásadním odchylkám, každý text anotují zároveň dva anotátoři, jejichž výsledek se poté podrobí srovnání, adjudikaci (Šebesta a Škodová 2012, s. 77). Při adjudikaci se z dvou paralelně zpracovaných textů vytvoří výsledný text, na kterém poté ještě proběhne automatická anotace[28]. Po jejím dokončení je text připraven na zařazení do korpusu.

Obrázek 9: Náhled na anotační program *feat* spolu s rozhraním korpusového manažeru *Speed* (panel vlevo)



[27] K problému mezianotátorské shody viz disertační práci B. Štindlové (Štindlová 2011, s. 121).

[28] Nutno poznamenat, že v nynější fázi zpracování korpusu nebyla adjudikace ve větší míře dosud provedena.

5 Popis korpusu FALKO

Chybový korpus němčiny jako cizího jazyka[29] (neboli FALKO – Fehlerannotiertes Lernerkorpus) je projektem Humboldtovy univerzity v Berlíně. Jedná se o korpus psané němčiny jako cílového jazyka, výchozí jazyk mluvčích je různý.

V době před vznikem korpusu FALKO existovalo jen málo korpusů němčiny jako cílového jazyka, kromě toho byly velmi malé a většina z nich nebyla veřejně přístupná (Lüdeling et al. 2006, s. 1). Vybavenost chybovou anotací u těchto korpusů téměř neexistovala. Jediný korpus, který byl chybově anotován, zpracovala Ursula Weinberger z Lancasterské univerzity, bohužel i tento korpus byl velmi malý (95 textů, 27635 slov) a veřejně nepřístupný (Lüdeling et al. 2006, s. 1).

Korpus FALKO tedy vznikl z potřeby vytvořit přístupnou platformu pro výzkum němčiny jako cizího jazyka. Velmi obecný účel korpusu, neomezující se pouze na určitý výzkumný záměr, předpokládá takovou korpusovou architekturu, která bude schopná zachytit v co největší míře specifika zkoumaného mezijazyka a bude dostatečně flexibilní, aby zajistila znovupoužitelnost korpusu. Z tohoto důvodu byla uvedena korpusová architektura založená na vícerovinném, distančním anotačním schématu. Výhody tohoto přístupu popsali autoři například v příspěvku *Multi-level error annotation in learner corpora* z konference o korpusové lingvistice Birmingham 2005 (Lüdeling et al. 2006) či nověji v článku *Competing Target Hypotheses in the Falko Corpus. A Flexible Multi-Layer Corpus Architecture* (Reznicek et al. v přípravě). Teze, které autoři v článcích prezentují, budou představeny v kapitole 5.3 této práce.

Korpus FALKO se skládá z několika subkorpusů. Odlišujícím faktorem je typ úlohy, při které jazyková data vznikají, a jazyk (Reznicek et al. 2012).

Na základě typu úlohy se FALKO dělí na subkorpus obsahů/shrnutí (FalkoSummary) a subkorpus esejí (FalkoEssay).[30] Subkorpus FalkoSummary je trojího typu, podle jazyka produkce – FalkoSummaryL2, který obsahuje texty psané mluvčími němčiny jako cizího jazyka, FalkoSummaryL1, což je kontrolní korpus textů stejného zadání psaných

[29] Viz webové stránky korpusu FALKO (Dietterle 2013c).

[30] Tento a následující dva odstavce přebírají informace z příručky *Das Falko-Handbuch* (Reznicek et al. 2012, s. 4).

německými rodilými mluvčími a FalkoSummaryVL, který se skládá z originálních textů, na jejichž základě se obsahy/shrnutí psaly.

Podobné rozdělení platí i pro subkorpus FalkoEssay s tou výjimkou, že samotný subkorpus cizojazyčných mluvčích je rozdělen na dvě části, FalkoEssayL2 a FalkoEssayL2WHIG, a neexistuje subkorpus originálních textů (zadání úlohy je nevyužívá). Část FalkoEssayL2WHIG má zvláštní označení, protože je součástí projektu WHiG („What's hard in German"), řešeným taktéž Humboldtovou univerzitou, viz stránky projektu (Suppus 2013). Projekt má za cíl „[...] systematicky identifikovat lingvistické struktury, které představují zvláštní potíže pro akvizici němčiny jako cizího jazyka [...]" (Suppus 2013, přeloženo), k čemuž slouží taktéž právě korpus FALKO.

Specifické postavení má subkorpus FalkoGeorgetown, který je longitudinálního charakteru. Subkorpus je dále dělen na základě jazyka na FalkoGeorgetownL2 (cizojazyční mluvčí němčiny) a FalkoGeorgetownL1 (srovnávací korpus rodilých mluvčích).

Schéma rozdělení korpusu FALKO spolu s velikostmi jednotlivých části ukazuje následující tabulka:

Tabulka 3: Schéma korpusu FALKO[31]

-	Korpus cizojazyčných mluvčích	Srovnávací korpus rodilých mluvčích	Korpus originálních textů	∑
Korpus obsahů/ shrnutí	FalkoSummaryL2v1.2 (40 638 tokenů)	FalkoSummaryL1v1.2 (21 211 tokenů)	FalkoSummaryVL (11 016 tokenů)	72 865
Korpus esejí	FalkoEssayL2v2.4 (144 619 tokenů)	FalkoEssayL1v2.3 (70 615 tokenů)		215 234
	FalkoEssayL2WHIGv2.0 (130 187 tokenů)			130 187
Longitudinální korpus	FalkoGeorgetownL2v1.0 (125 993 tokenů)	FalkoGeorgetownL1v1.0 (12 668 tokenů)		138 661
∑	441 437	104 494	11 016	556 947

V tabulce 3 jsou data z dotazovacího systému *ANNIS3* (ANON. 2013e) a z

[31] Podle *Falko-Handbuch* (Reznicek et al. 2012, s. 4). Přeloženo a upraveno tak, aby byl reflektován nejnovější stav (přidán korpus FalkoGeorgetown a aktualizovány údaje o velikosti podle dotazovacího systému ANNIS3(ANON. 2013e).

dokumentace subkorpusu FalkoGeorgetown (Lüdeling 2007). Aktuální celková velikost v tokenech je 556 847. Referenční příručka *Falko-Handbuch* (Reznicek et al. 2012) z roku 2012 uvádí sumu menší: 381 447. Údaje v příručce totiž nereflektují hodnoty longitudinálního subkorpusu (Dietterle 2013b).

Ani suma v tabulce však není konečná, počítá se s postupným přidáváním dalších jazykových dat, korpus je tzv. „ever-growing" (Dietterle 2013b).

Celý korpus i s metadaty je k dispozici ke stažení na stránkách katedry německého jazyka a lingvistiky Humboldtovy univerzity v Berlíně, viz (Dietterle 2013a), a je možné ho prohledávat v systému *ANNIS3*, který je taktéž umístěn na webových stránkách Humboldtovy univerzity, viz (ANON. 2013e).

5.1 Jazyková data korpusu

FALKO je korpus pouze písemný. Skládá se z textů, které byly klinicky elicitovány ve výukovém kontextu. Jak už bylo řečeno výše, jednotlivé subkorpusy se liší typem úlohy, kterou mluvčí vykonávali.

5.1.1 FalkoSummary [32]

Jazyková data sukorpusu FalkoSummaryL2 pochází od studentů germanistiky na Freie Universität Berlin (Reznicek et al. 2012, s. 11). Každý student musí v určité fázi studia úspěšně složit test, který ověří jeho znalosti německého jazyka na úrovni porozumění a produkce odborného textu. Studenti obdrží zadání s odborným textem, zaměřeným na německou lingvistiku či literaturu, jehož obsah musí v limitu 90 minut shrnout a popsat vlastními slovy. Studenti nemají texty předem k přípravě a nemohou používat žádných pomůcek. Všichni studenti, skládající tento test, jsou pokročilí mluvčí (absolvovali DSH-test, který je ekvivalentní úrovni C1-C2 evropského referenčního rámce pro jazyky).

K FalkoSummaryL2 se váže srovnávací subkorpus FalkoSummaryL1, který obsahuje texty psané na základě stejných zadání rodilými mluvčími němčiny, studenty z Freie Universität Berlin a Humboldtovy univerzity. Podmínky byly stejné jako u subkorpusu

[32] Kapitola zpracována podle FALKO *Summary Documentation* (Jia Wei 2007, s. 1).

nerodilých mluvčích s tou výjimkou, že se nejednalo o zkoušku.

Celek korpusu uzavírá FalkoSummaryVL (VL=Vorlagen, „něm. předloha, zadání“), což je korpus originálních textů, které byly součástí zadání. Jejich seznam je k nalezení ve Falko Summary Documentation, viz (Jia Wei 2007, s. 11).

5.1.2 FalkoEssay[33]

Texty subkorpusu FalkoEssayL2 jsou argumentativní eseje nerodilých mluvčích němčiny na předem daná témata Sběru se účastnila Humboldtova univerzita, Freie Universität Berlin a částečně také Goethe-instituty v zahraničí a některé zahraniční univerzity. Podmínky elicitace byly na všech místech stejné: žádné předchozí seznámení s textem, žádné pomůcky, čas na vypracování 90 minut. Jazyková kompetence mluvčích byla určena pomocí C-testu[34] a kolísá od úrovně B2 až po úroveň C2.

Stejně koncipovaný je i druhý subkorpus FalkoEssayL2WHIG. Jako součást projektu WHiG však byly texty shromážděny jinými subjekty. Jedná se o partnerský projekt Humboldtovy univerzity a velšské univerzity Bangor. Mimo univerzitu Bangor se na sběru textů podílí i další britské univerzity, viz (Reznicek et al. 2012, s. 23). FalkoEssayWHIG je tedy primárně zaměřen na anglické studenty němčiny jako cizího jazyka.

Srovnávací subkorpus FalkoEssayL1 obsahuje produkci středoškolských studentů, rodilých mluvčích němčiny. Texty byly rovněž vypracovány za stejných podmínek a byly stejného zadání, jako texty nerodilých mluvčích.

5.1.3 FalkoGeorgetown

Jazyková data korpusu FalkoGeorgetownL2 se sbírala na americké Georgetown University ve Washingtonu, D.C. (Lüdeling 2007, s. 1). Studenti, kteří na této univerzitě studovali němčinu jako cizí jazyk, napsali v průběhu svého studia (zpravidla na konci jednoho úseku) celkem čtyři texty, reflektující jejich jazykový vývoj, takzvané *PPTs* (Prototypical Writing Tasks) (Lüdeling 2007, s. 2). Jednalo se o dopis, vyprávění, novinový článek a proslov (Lüdeling 2007, s. 2–5). Mimo tyto prototypické texty vypracovali studenti také recenze na knihy. Tato úloha se označuje jako *BWTs* (Baseline Writing Tasks)

[33] Informace v této kapitole pochází z *Falko-Handbuch* (Reznicek et al. 2012, s. 23–24).

[34] Viz http://www.c-test.de/ (ANON. 2012).

(Lüdeling 2007, s. 2).

Ne všechny texty však byly zařazeny do vyhledávatelného korpusu. Longitudinální subkorpus FalkoGeorgetownL2, dostupný skrze vyhledávací rozhraní *ANNIS3*, je tvořen pouze prototypickými texty (*PPTs*) o celkové velikosti 78 132 tokenů .

Srovnávací subkorpus FalkoGeorgetownL1 je složen z textů rodilých německých mluvčích. Jedná se o knižní recenze, které napsali studenti a učitelé univerzit Freien Universität Berlin, Georgetown University a univerzity Trier, část textů byla taktéž pořízena na internetu, viz (Lüdeling 2007, s. 15).

Pokud srovnáme informace z této kapitoly s popisem korpusu CZESL, vyvstanou nám zde jasné odlišnosti. Korpus FALKO obsahuje texty žánrově omezené s předem danými tématy, zaměřuje se z větší části pouze na pokročilé mluvčí a jeho součástí je jasně vymezený longitudinální subkorpus.

5.2 Metadata

Metadata korpusu FALKO pokrývají základní údaje o textu, autorovi a jeho jazykové kompetenci (Reznicek et al. 2012, s. 9–11). Vzhledem k tomu, že je FALKO korpusem pouze psané řeči, systém metadat je jednodušší, než v případě CZESLu. Tabulka X ukazuje metadata subkorpusů FalkoSummary, FalkoEssay a FalkoGeorgetown.

Tabulka 4: Metadata korpusu FALKO[35]

Parametry spojené s textem	
FalkoEssay a FalkoSummary	FalkoGeorgetown
corpus (název korpusu)	corpus (název korpusu)
subcorpus (označení subkorpusu)	transcriptionName (označení přepisu)
transcriptionName (označení přepisu)	exercise (typ cvičení = PPTs/BWTs)
collectionDate (datum sběru)	level (úroveň úlohy PPTs/BWTs)
corrector (autor cílové hypotézy)	
originalFilename (originální název souboru)	
topic (téma úlohy)	

[35] Zpracováno podle *Falko-Handbuch* (Reznicek et al. 2012, s. 9–11) a souborů se souhrnem korpusových metadat na stránkách projektu (Dietterle 2013c).

Parametry spojené s autorem	
FalkoEssay a FalkoSummary	FalkoGeorgetown
name (zašifrované příjmení)	learner_id (id žáka)
firstName (zašifrované jméno)	semester (semestr studia)
birthYear (rok narození)	
sex (pohlaví)	
mayorSubject (předmět)	
degree (dosažené vzdělání)	
ctest (výsledek C-testu v *n*)	
Parametry spojené s jazykovou kompetencí autora	
FalkoEssay a FalkoSummary	
l1_*n* (*n*-tý mateřský jazyk)	
l1_*n*_since (odkdy se učí jazyk; u mateřského 0)	
l1_*n*_duration (jak dlouho užívá jazyk, v měsících)	
l1_*n*_school (učil se jazyk ve škole?)	
l1_*n*_university (učil se jazyk na univerzitě?)	
l1_*n*_langschool(učil se jazyk v jazykové škole?)	
l1_*n*_awaymonths (pobyt v zemi cílového jazyka, v měsících)	
l1_*n*_awayplace (místo pobytu v zemi cílového jazyka; u mateřského jazyka místo, kde se v raném dětství učil mluvit)	
l2_*n* (*n*-tý cizí jazyk; první cizí jazyk v pořadí ovládá nejlépe, druhý hůře atd.)	
l2_*n*_since (odkdy se učí jazyk)	
l2_*n*_duration (jak dlouho užívá jazyk, v měsících)	
l2_*n*_school (učil se jazyk ve škole?)	
l2_*n*_university (učil se jazyk na univerzitě?)	
l2_*n*_langschool(učil se jazyk v jazykové škole?)	
l2_*n*_awaymonths (pobyt v zemi cílového jazyka, v měsících)	
l2_*n*_awayplace (místo pobytu v zemi cílového jazyka)	

Jak vidíme v tabulce 4, FalkoGeorgetown užívá v kontrastu s ostatními subkorpusy pouze základní, nevelké množství metadat.

Protože jsou podmínky vzniku a sběru textů u jednotlivých subkorpusů identické, neuvádí se podrobnější metadata spojená s touto situací (narozdíl od korpusu CZESL, kde je tato oblast zpracována velmi podrobně). Podrobné jsou však metadata mapující jazykovou vybavenost autora, dle užitého rámce se počítá se systematickým záznamem úplné jazykové kompetence autora.

5.3 Anotace

Stěžejní prvkem korpusu FALKO je vícerovinný anotační model. Zavedení tohoto systému je obzvláště vhodné pro žákovské korpusy či jiné korpusy, které se potýkají s výraznými odchylkami od „standardního“ jazyka (Reznicek et al. v přípravě, s. 1). Vícerovinná anotace totiž představuje možnost, jak eliminovat nedostatky tzv. lineárního anotačního modelu (flat annotation model)[36] – anotaci disparátních jednotek, překrývající se anotaci a problém cílové hypotézy (Reznicek et al. v přípravě, s. 2–6).

Pokud se podíváme na tabulární model lineární anotace, vidíme, že je zde nemožné anotovat více jednotek najednou, viz tabulka 5 a 6.

Tabulka 5: Příklad POS tagů a lemmatizace v tabulárním anotačním modelu[37]

I/PP/I **have**/VBP/have **got**/VBN/get **,**/,/, **Monday**/NP/Monday **or**/CC/or **Tuesday**/NP/Tuesday **off**/RB/off

Tabulka 6: Příklad chybové anotace v tabulárním anotačním modelu[38]

<LxPhCh> Es gibt eine veränderte Gesellschaft und...

Překlad: There is a changed society and... (Existuje změněná společnost...)

Lx – lexikum, Ph – fráze, Ch – nesprávný výběr

Morfologický tag a lemma jsou přiřazovány na úrovni tokenu přímo do textu, pro každý token zvlášť, což znemožňuje jejich spojování (Lüdeling et al. 2006, s. 5). U chybové anotace chybí explicitní cílová hypotéza a neexistuje zde označení, jakých tokenů se chybový tag týká.

Částečně tyto problémy řeší stromový systém anotace, viz tabulka 7.

[36] Označuje se taktéž jako vkládaná anotace (inline mark-up).

[37] Převzato z (Lüdeling et al. 2006, s. 4), zvýraznění provedl autor této práce.

[38] Převzat z (Lüdeling et al. 2006, s. 5), zvýraznění pochází od autorů odkazovaného článku; přeloženo.

Tabulka 7: Stromová lineární anotace[39]

L'héritage du passé est très <G><GEN><ADJ> #fort$ forte </ADJ></GEN></G> et le sexisme est toujours présent.

Překlad: The heritage of the past is very strong and the sexism is always present.

(Dědictví minulosti je velmi *silná a sexismus je stále přítomen.)

G – typ chybové domény, GEN – subdoména, ADJ – určení slovního druhu

Ten umožňuje jasně vymezit působnost chybové značky, zároveň je možné explicitně vyjádřit cílovou hypotézu (v příkladu předchází chybnému tvaru „forte“). Problém však nastává, pokud budeme chtít anotovat token, který už byl jednou anotován v rámci jiné chyby. Stromová struktura se tímto krokem rozpadá.

Ani jeden z typů vkládané anotace však není schopen ošetřit případ konkurenčních cílových hypotéz.

V článku *Multi-level error annotation in learner corpora* (Lüdeling et al. 2006, s. 3) uvádějí autoři následující situaci, viz tabulka 8.

Tabulka 8: konkureční cílové hypotézy[40]

die Erklärung für **<MoArInGn>diese** Phänomen ist einfach

Překlad: the explanation for these phenomenon is simply (vysvětlení pro *tuto fenomén je jednoduché)

(Mo – morgolofie, Ar – člen, In – flexe, Gn – rod)

U příkladu se nabízí dvě cílové hypotézy: chyba v rodě u zájmena (*diese > dieses), jak ji upřednostnil anotátor, a chyba v čísle u podstatného jména (*Phänomen > Phänomene). Při lineární anotaci se musí anotátor vždy rozhodnout pouze pro jednu z konkurenčních hypotéz, ačkoli obě mohou mít stejnou váhu (jako v příkladu v tabulce 8).

Vícerovinná architektura korpusu FALKO tyto problémy řeší zavedením

[39] Převzato z (Lüdeling et al. 2006, s. 5), přeloženo.

[40] Převzato z (Lüdeling et al. 2006, s. 2), zvýraznění pochází od autorů odkazovaného článku; přeloženo.

pohyblivého počtu anotačních rovin, kdy na každé rovině může stát jedna cílová hypotéza, a distanční anotací, kdy značky nejsou vkládány přímo k tokenům, ale jsou vytvářeny mimo ně a pouze na ně odkazují.

5.3.1 Chybová anotace

Anotační formát korpusu FALKO má charakter tabulky. Řádky tabulky představují jednotlivé anotační roviny, jednotlivá pozice na řádku označuje tentýž token na všech rovinách, viz tabulku 9. Tímto způsobem se mezi tokeny udržují jasné korespondence. Tabulační formát také umožňuje tokeny spojovat či rozpojovat, viz znova tabulku 9.

Tabulka 9: FALKO jako tabulka

Rovina 1	pozice a	pozice b+c		pozice d	pozice e	pozice f
Rovina 2	pozice a	pozice b	pozice c	pozice d	pozice e+f	

Způsob anotace jakožto i počet anotačních rovin se u jednotlivých subkorpusů liší, u všech se však minimálně provádí morfologická anotace (rovina „POS“) a lemmatizace (rovina „lemma“), viz tabulka 10 (Reznicek et al. 2012, s. 4)

Tabulka 10: POS a lemmatizace[41]

word	möchtet	(	wie	üblicherweise	die	meisten	)	braucht	man	meiner
pos	VMFIN	$(	KOKOM	ADV	ART	PIAT	$(	VVFIN	PIS	PPOSAT
lemma	mögen	(	wie	üblicherweise	d	meist	)	brauchen	man	mein

První rovina označená jako „word“ je přepis originálního žákovského textu rozdělený na jednotlivé tokeny. Roviny „lemma“ a „pos“ jsou výstupy automatické lemmatizace a automatického morfologického značkování – děje se tak pomocí programu Treetagger, viz (Schmid 2013). Více o lingvistické anotaci viz kapitolu 5.3.2 na straně 41 této práce.

[41] Text příkladu pochází z korpusu FalkoEssayL2v2.4, dokument cbs001_2006_09_L2v2.4.

5.3.1.1 FalkoEssay

Kromě již zmíněné lingvistické anotace (POS a lemma) jsou subkorpusy FalkoEssay a FalkoEssayWHIG anotovány minimálně dvěma cílovými hypotézami a dalšími doplňujícími rovinami. Přehled kompletní anotace je spolu s vysvětlivkami i příklady k nalezení ve *Falko-Handbuch* (Reznicek et al. 2012, s. 6–8), my zde uvádíme pouze zjednodušenou verzi, která však pro naše potřeby dostačuje, viz příloha 2, s. 66 této práce.

Jádrem anotace jsou dvě cílové hypotézy, tzv. minimální (TH1) a maximální (TH2). Rozdíly mezi nimi popisuje Reznicek et al. (v přípravě, s. 16–17, přeloženo):

- Minimální cílová hypotéza:
  A. minimální gramatické úpravy, operuje na větné úrovni
  B. výsledná cílová hypotéza je gramaticky správná

  Výhody:
  1. relativně jednoduchá anotační pravidla
  2. vysoká mezianotátorská shoda
  3. je strukturně blízká původnímu textu

  Nevýhody:
  1. stále může obsahovat chyby
- Maximální cílová hypotéza:
  A. využívá sémantiky a pragmatiky, operuje na textové úrovni
  B. výsledná cílová hypotéza je gramaticky správná, navíc sémanticky koherentní a pragmaticky přijatelná

  Výhody:
  1. přibližuje se předpokládanému záměru mluvčího
  2. zahrnuje lingvistické informace ‚vyšší úrovně'

  Nevýhody
  1. je otevřená množství rozdílných interpretací
  2. může vést ke značným změnám v povrchové struktuře textu

Zjednodušeně řečeno, minimální hypotéza se snaží co nejméně zasahovat do původního textu (zahrnuje pravopis, morfosyntax), zatímco maximální cílová hypotéza usiluje nejen o gramatickou správnost, ale i o opravy ku prospěchu srozumitelnosti

a smyslu textu (zahrnuje sémantiku, pragmatiku, lexikum).

Příklad takovéto anotace zobrazuje tabulka 11:

Tabulka 11: Anotace cílových hypotéz[42]

tok	Hráč	vstřelil	na	branku	a	všichni	se	smáli	.	1	:	0	.
TH1	Hráč	vystřelil	na	branku	a	všichni	se	smáli	.	1	:	0	.
TH1Diff		CHANGE											
TH2	Hráč	vstřelil		gól	a	všichni	se	radovali	.	1	:	0	.
TH2Diff			DEL	CHANGE				CHANGE					

Každá cílová hypotéza (zde TH1 a TH2) je explicitně vyjádřena na své vlastní rovině, k níž náleží podle potřeby další roviny, které tuto hypotézu popisují (viz kompletní schéma anotace, příloha 2, strana 66 této práce). V tomto zjednodušeném případě se jedná o roviny TH1Diff a TH2Diff, kde je chyba vymezena a opatřena chybovým tagem. Chybové tagy na rovinách „Diff" označují změny v povrchové relizaci textu, tedy změnu tokenu, jeho přesun, vložení či vymazání. Taxonomii chyb povrchové realizace zachycuje tabulka 12:

Tabulka 12: taxonomie chyb povrchové realizace[43]

Tag	Popis
INS	vložení tokenu na TH
DEL	smazání tokenu na TH
CHA	změněný token na TH
MOVS	zdrojová lokace přesunutého tokenu na TH
MOVT	cílová lokace přesunutého tokenu na TH
MERGE	tokeny spojeny na TH
SPLIT	tokeny rozděleny na TH

V příkladu anotace cílových hypotéz (tabulka 11) je zobrazeno smazání předložky „na" – důležité je, že pozice v tabulce zůstává jako prázdná buňka, nevymazává se. Podobně je to i u ostatních operací, viz následující tabulky:

[42] Podle (Reznicek et al. v přípravě, s. 21–23); příkladová věta pochází od autora této práce.

[43] Podle (Reznicek et al. v přípravě, s. 21), přeloženo.

Tabulka 13: příklad opravy slovosledu

ctok	Hráč	vystřelil	na	branku	a	se	všichni		smáli
TH1	Hráč	vystřelil	na	branku	a		všichni	se	smáli
TH1Diff						MOVS		MOVT	

Tabulka 14: příklad opravy spojením a rozpojením

ctok	Hráč	vystřelilna		branku	a	všich	ni	se	smáli
TH1	Hráč	vystřelil	na	branku	a	všichni		se	smáli
TH1Diff		SPLIT				MERGE			

Kromě anotace chyb povrchové realizace probíhá ještě chybová anotace německých komplexních sloves, což jsou slovesa složená z oddělitelné či neoddělitelné části a slovesné části, například „be-kommen“, „an-rufen“), viz (ANON. 2010). Slovesa s oddělitenou částí (což se projevuje například při konjugaci) jsou v korpusu nazývána „Patrikelverb“, neoddělitelná „Präfixverb“. Kvůli nejasné definici považují autoři za komplexní slovesa i tvary složené z infinitivu („kennenlernen“) anebo jména („klavierspielen“) (Reznicek et al. 2012, s. 63).

Tento typ sloves má v korpusu svou rovinu cílové hypotézy (*THverb*) a kompletní lingvistickou anotaci. Autoři chtěli shromáždit co nejvíce jevů vázající se na komplexní slovesa, proto byly také zavedeny speciální anotační roviny, na kterých se rozlišuje kategorie těchto sloves (*verbkategorie*), jejich lemma (*verblemma*) a slovesný tvar (*verbform*) (Reznicek et al. 2012, s. 63). Pokud je přítomna chyba, anotuje se na rovině *verbfehlertyp*.

Anotují se chyby lexikání (chybový tag *sem*, *lex*), ortografické (*orth*), chyby v nadužití neverbální části komplexního slovesa (*part*), argumentační struktuře (*as*), chyby ve flexi (*flex*) a slovosledu (*ws*) (Reznicek et al. 2012, s. 64–65). Pod chyby ve flexi patří taktéž chybové kódy *ge* a *zu*, které označují špatné užití těchto předpon (chybějící „ge“ či jeho užití na špatném místě apod.). Pro specifičtější chybu ve špatném rozdělení komplexního slovesa se taktéž zavedl speciální tag – *sep* (patří pod chybový kód *ws*).

Úplný seznam anotačních kódů je v příloze 3 na straně 68 této práce, podrobný výpis

i s příklady je k nalezení ve zdrojové publikaci (Reznicek et al. 2012, s. 63–66).

Poslední anotační kategorií v korpusu FalkoEssay je závislostní analýza (roviny *dep* a *func*), pomocí níž lze konstruovat tzv. „Dependenzenbaüme“ (závislostní stromy) (Reznicek et al. 2012, s. 62). Závislostní analýza probíhá pomocí automatických nástrojů na základě předchozí anotace (POS s následnou ruční korekcí).

5.3.1.2 FalkoSummary

Chybová anotace sukorpusu FalkoSummaryL2 je založena na formulaci cílových hypotéz podle 8 lingvistických kategorií. Jedná se o pravopis, slovotvorbu, shodu, dominaci, čas, způsob, slovosled a styl/lexikum (Lüdeling et al. 2006, s. 7). Jednotlivé cílové hypotézy – roviny, se mohou přidávat podle potřeby. Jsou anotovány ve třech stupních odpovídajících krokům chybové analýzy: identifikace, popis a explanace (Šebesta a Škodová 2012, s. 53). Pro názornou ukázku ilustrujeme toto anotační schéma tabulkou 15, převzatou z článku *Multi-level error annotation in learner corpora* (viz Lüdeling et al. 2006, s. 9):

Tabulka 15: Příklad anotace cílové hypotézy korpusu FalkoSummaryL

word	dass	nur	er	…	konnte	durch	dieses	Tor	eingelassen	werden	
target					durch dieses Tor eingelassen werden konnte						
word order identification					**x**						
word order description					MF_RSK						
word order explanation					transfer						

První rovinu v tabulce zaujímá originální segmentovaný text, pod ním je samotná cílová hypotéza, která se v tomto případě týká opravy slovosledu (word order). Rovina identifikace (identification) graficky vymezuje působnost opravy. Rovina popisu (descripion) je založena na teorii větné struktury v němčině, která rozděluje větu na určité topologické celky, tzv „Vorfeld“ (VF), „Mittelfeld“ MF) a „Nachfeld“ (NF), neboli počáteční pole, střední pole a konečné pole, a taktéž „Linke Satzklammer“ (LSK) a „Rechte Satzklammer“ (RSK) (Lüdeling et al. 2006, s. 8). Chybový tag indikuje špatné umístění v „Mittelfeld“, středním poli, namísto „Rechte Satzklammer“, pravé části věty.

Rovina explanační pak už jen popisuje chybu jako přesun.

Kromě automatické lingvistické anotace (lemmatizace a POS) a formulace cílové hypotézy je součástí anotace korpusu FalkoSummaryL2 taktéž popis již zmíněných topologických polí („Topologische Felder“) a syntaktický popis („Syntaktische Beschreibung“). Celkové schéma anotace je přiloženo jako příloha 4, s. 69 této práce.

5.3.2 Lingvistická anotace

Lingvistická anotace v korpusu FALKO je dvojího druhu – POS a lemmatizace. Obojí probíhá u každého korpusu automaticky pomocí nástroje *Treetagger*[44] (Reznicek et al. 2012, s. 4).

POS anotace používá tagset nazvaný The Stuttgart-Tübingen Tagset (TSST)[45], který obsahuje 54 značek. Jejich seznam je uveden v příloze 5 na straně 70 této práce.

Paralelně s anotací *Treetaggeru* probíhá u korpusu FalkoEssayWHIG ještě alternativní lingvistická anotace pomocí programu *rfTagger*[46], který přiřazuje POS a rozšiřující morfologické tagy (roviny *rfPOS*, *rfMorph* apod.) (Reznicek et al. 2012, s. 4).

Kromě automatické anotace se provádí i manuální lingvistická anotace u subkorpusu FalkoSummary (rovina *cpos*).

5.4 Průběh anotace

Anotace korpusu FALKO má následující průběh (podle Reznicek et al. 2012, s. 4–5):

- žákovské texty se nejprve lingvisticky anotují
- poté postupují k manuální anotaci cílových hypotéz
- na základě předešlých anotací se anotují zbylé roviny
- kompletně anotovaný text se vkládá do korpusového vyhledávače *ANNIS* (ANON. 2013e), nyní ve verzi 3, viz obrázek 12.

Cílové hypotézy se nejprve zpracovávaly v programu *EXMARaLDA*[47] (viz obr. 10), ten byl však později vystřídán zásuvným modulem do aplikace MS Excel[48] (viz obr. 11).

---

[44] Viz (Schmid 2013).

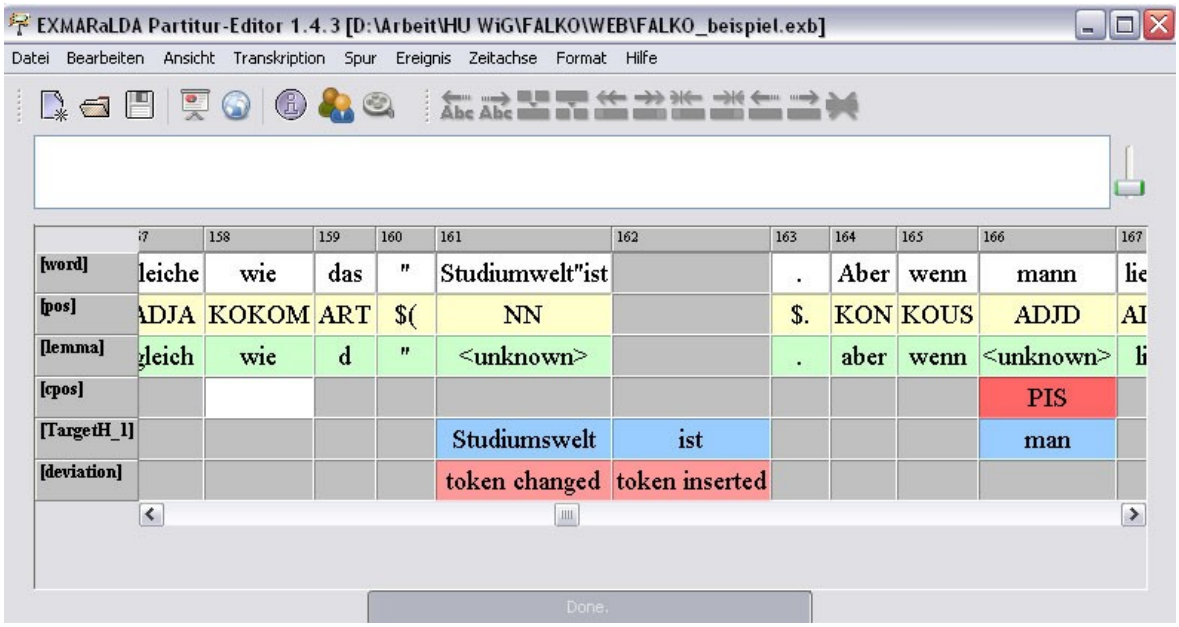
[45] Podrobnější informace viz stránky The Stuttgart-Tübingen Tagset (ANON. 2003).

[46] Viz (Schmid a Laws 2008).

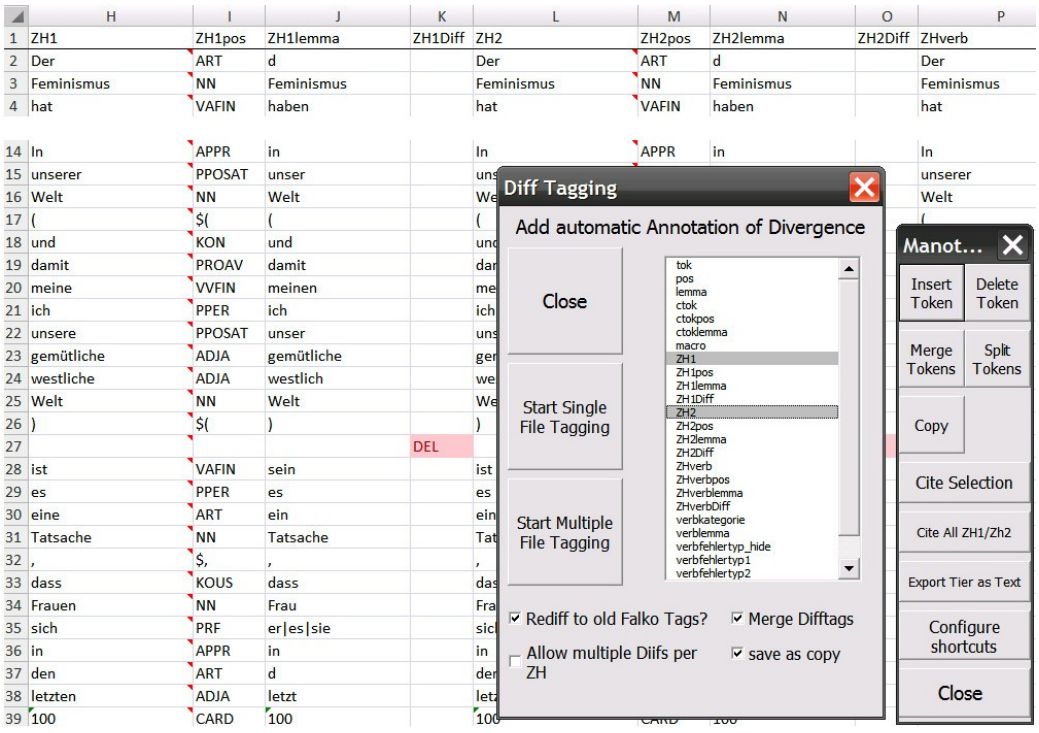
[47] (ANON. 2011); viz webové stránky www.exmaralda.org.

[48] (Zeldes 2013); viz webové stránky http://www.exmaralda.org/exceladdin.html.

Obrázek 10: Anotace v programu *EXMARaLDA*[49]



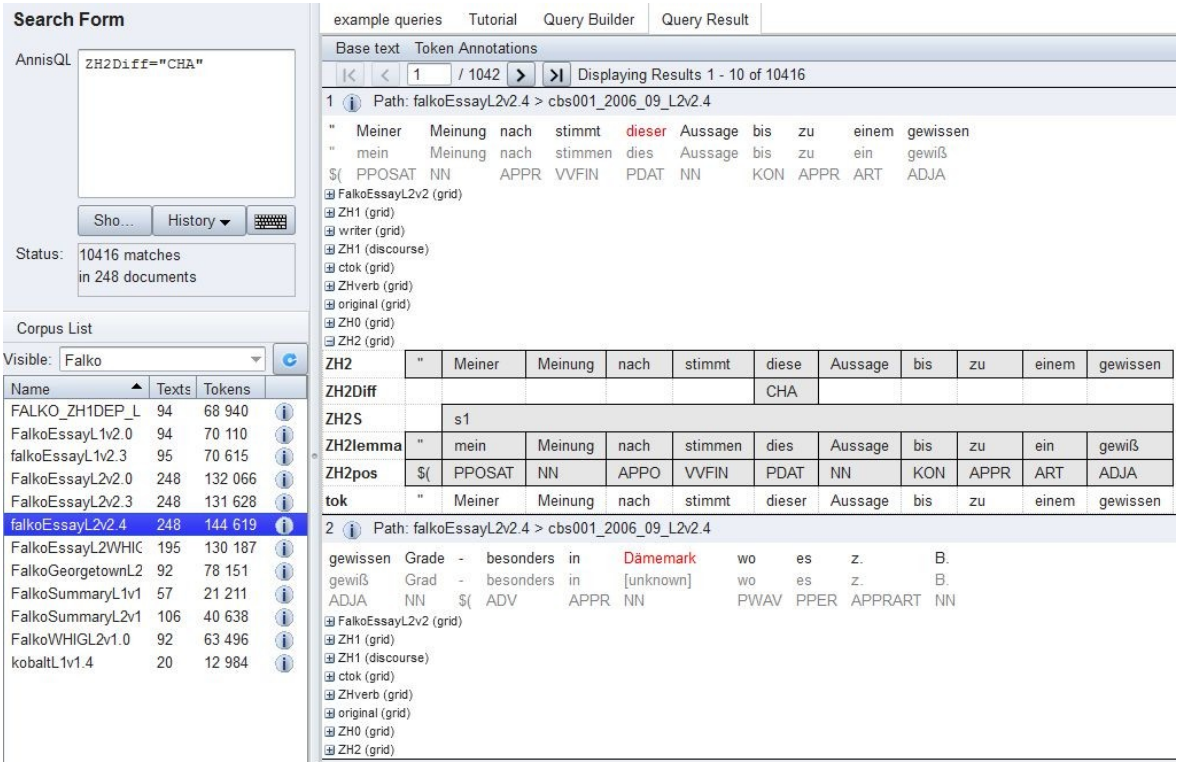
Obrázek 11: Anotace v zásuvném modulu pro MS Excel[50]



[49] Převzato ze stránek projektu exmaralda.org.

[50] Převzato ze stránek projektu exmaralda.org, upraveno.

Obrázek 12: Rozhraní *ANNIS3*[51]



[51] Snímek obrazovky internetového prohlížeče při spuštěném rozhraní *ANNIS3* na adrese https://korpling.german.hu-berlin.de/falko-suche/.

6 Srovnání a exemplifikace

Z popisu jednotlivých korpusů v předešlých kapitolách vyplynulo několik zásadních rozdílů i podobností.

Oba korpusy jsou multilingválního charakteru, tedy zpracovávají texty, které jsou napsány množstvím různých jazyků. Korpus FALKO se však zaměřuje pouze na vyšší jazykové úrovně, zatímco korpus CZESL anotuje i texty začátečnické.

Podmínky sběru textů a jejich náplň jsou u korpusu FALKO standardizované, u korpusu CZESL jsou různorodé; závisí do jisté míry na spolupracujícím subjektu, který texty poskytuje. Korpus CZESL má z tohoto důvodu jemnější rozlišení metadat spojených se sběrem a vznikem textu.

Jedním ze základních požadavků anotačního systému korpusu CZESL je přiměřená náročnost anotace. Tagset korpusu CZESL je kompromisem mezi anotační nenáročností (a tím pádem i zjednodušení) a snahou o co nejpodrobnější zachycení jednotlivých chyb. Výhodou je větší mezianotátorská shoda. Jednoduchost anotace je naproti tomu ve Falku minoritní kritérium. Bere se jako výhoda (například je zmíněna při popisu minimální cílové hypotézy, viz Reznicek et al. v přípravě, s. 15–16), nicméně je zde tendence o co největší anotační možnosti a flexibilitu.

Hlavní rozdíl však spočívá v pojetí modelu chybové anotace. Jak vyplynulo z předchozích kapitol, oba korpusy užívají vícerovinné distanční anotace.

Korpus FALKO jako průkopník tohoto systému volí pohyblivý počet rovin, kterými se anotace může libovolně rozšiřovat. Rozdělení rovin je dáno buďto lingvistickými kritérii (FalkoSummary), anebo strukturně-procedurálními (FalkoEssay), v obou případech však platí, že lze anotační schéma dále upravovat.

Pojetí CZESLu je odlišné: předem byly určeny tři anotační roviny, přičemž první obsahuje originální text a druhá a třetí představuje jednotlivé fáze postupné opravy. Roviny postupné anotace jsou definovány lingvistickými kategoriemi.

Užité schéma předurčuje charakter chybové anotace. Ačkoli je u subkorpusu FalkoEssay zpracována syntaktická analýza i dependenční strom, návaznost této anotace na chybné tvary slov je pouze implicitní (výjimku tvoří anotace německých komplexních

sloves). Explicitně jsou vyjádřeny jen povrchové realizace chyb v rámci formulace cílových hypotéz.

Odlišná je situace u subkorpusu FalkoSummary, jehož anotace založená na procesu chybové analýzy a umožňuje do specifikace chyby zahrnout i bezprostřední kontext.

Charakter chybové anotace CZESLu se vymyká oběma popsaným subkorpusům. Explicitní morfologická či syntaktická definice je přítomna u každé chyby i s případným odkazem na formující prvek. CZESL tak umožňuje přímo vyznačit vliv okolních tokenů, zatímco u Falka jsou tyto vztahy naznačeny pouze implicitně.

Výše předložené poznatky budeme ilustrovat a doplňovat na základě reálných dat. Jak již bylo zmíněno v úvodních pasážích, výchozím korpusem pro srovnání bude korpus CZESL.

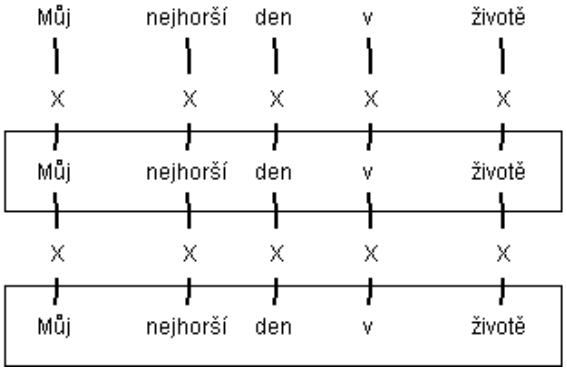
Text zpracovaný v rámci korpusu CZESL bude taktéž anotován v korpusu FALKO, aby se tak porovnaly možnosti jednotlivých anotačních modelů v praxi. Ač je korpus FALKO konstruován s ohledem na němčinu, jeho architektura je svým flexibilním formátem dosti univerzální, aby umožnila anotaci i jiného jazyka. Na následujících stranách bude provedena anotace části textu nejprve v korpusu CZESL a poté v korpusu FALKO.

6.1 CZESL

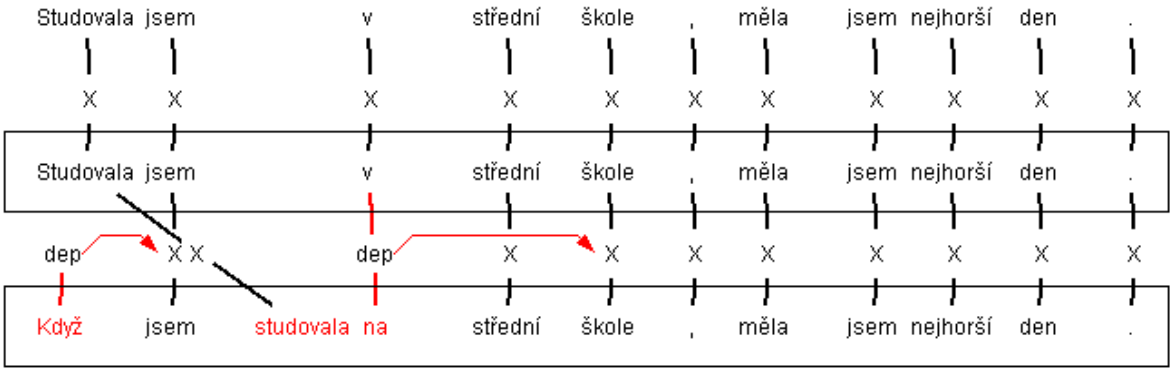
Anotace je provedena v programu *feat*. Výsledek anotace byl exportován zabudovaným nástrojem jako obrázek přímo z programu *feat*.

Text je kvůli své délce rozdělen na jednotlivé očíslované věty (souvětí).

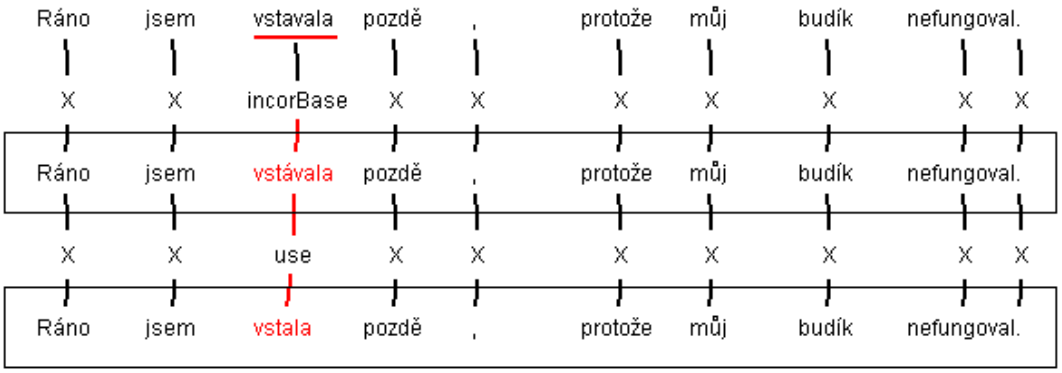
(1)



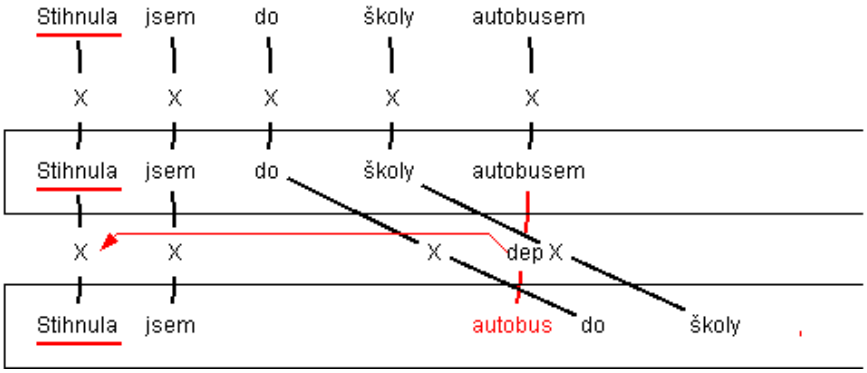
(2)



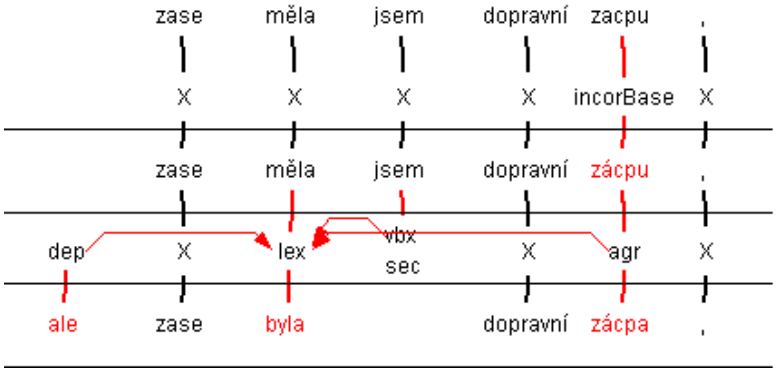
(3)



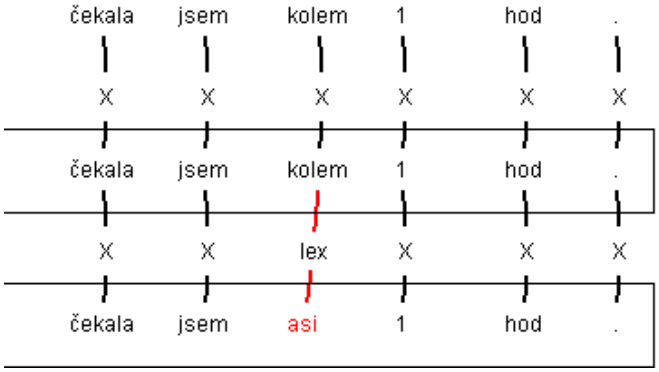
(4)


Stihnula jsem do školy autobusem
X X X X X
Stihnula jsem do školy autobusem
X X X dep X
Stihnula jsem autobus do školy ,

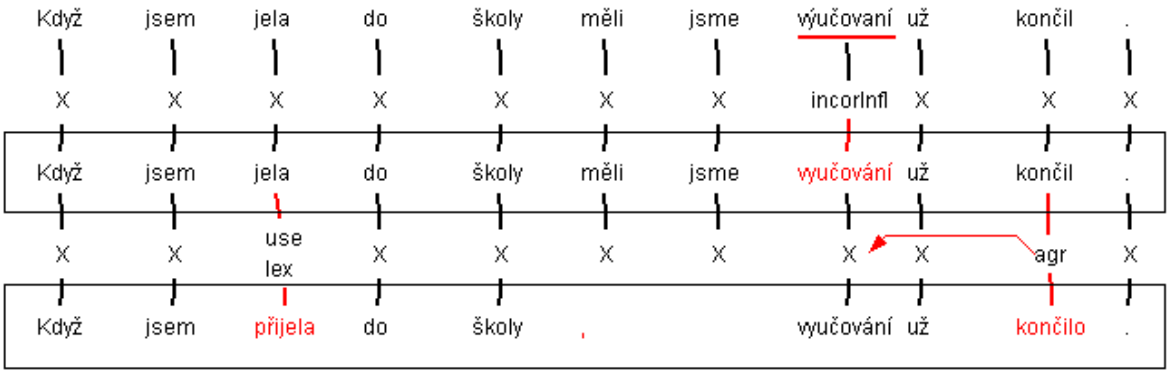
(5)


zase měla jsem dopravní zacpu ,
X X X X incorBase X
zase měla jsem dopravní zácpu ,
dep X lex vbx X agr X
sec
ale zase byla dopravní zácpa ,

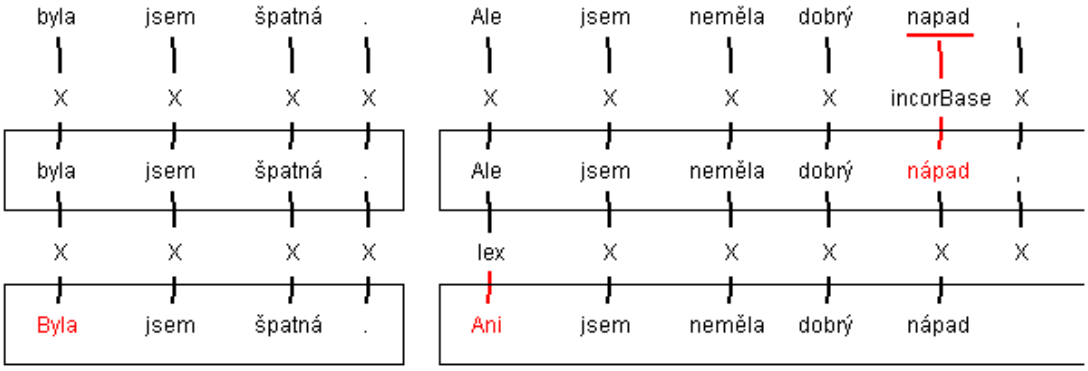
(6)


čekala jsem kolem 1 hod .
X X X X X X
čekala jsem kolem 1 hod .
X X lex X X X
čekala jsem asi 1 hod .

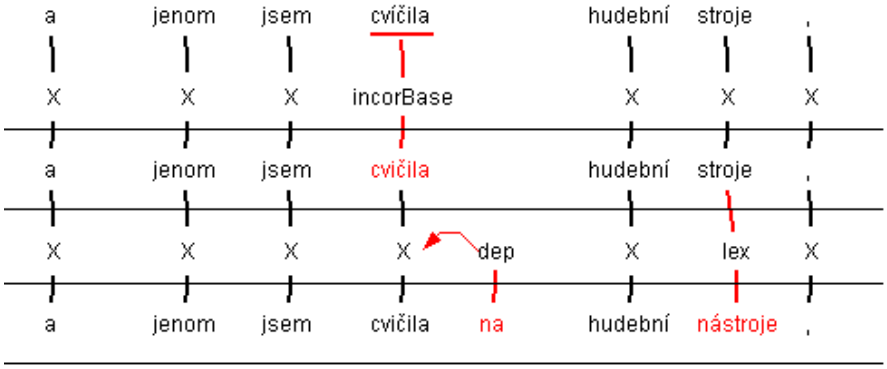
(7)


Když jsem jela do školy měli jsme výučovaní už končil .
X X X X X X X incorInfl X X X
Když jsem jela do školy měli jsme vyučování už končil .
X X use X X X X X X agr X
lex
Když jsem přijela do školy , vyučování už končilo .

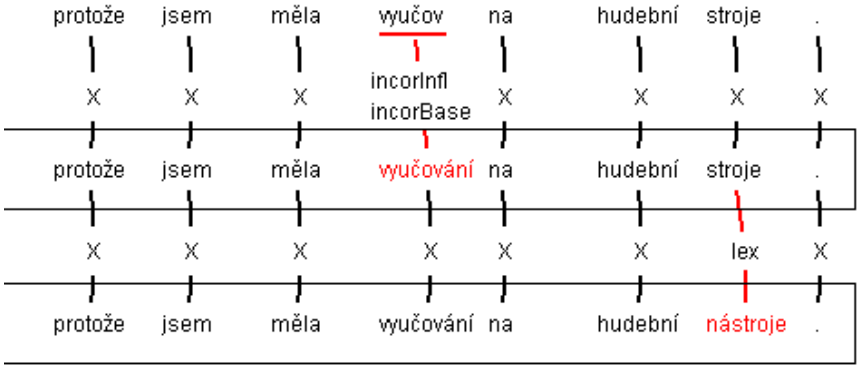
(8)


byla jsem špatná .
Ale jsem neměla dobrý napad ,
incorBase
nápad
lex
Byla
Ani

(9)


a jenom jsem cvičila hudební stroje ,
incorBase
dep
na
nástroje

(10)


protože jsem měla vyučov na hudební stroje .
incorInfl
incorBase
vyučování
lex
nástroje

6.2 FalkoEssay

Anotace je provedena v programu MS Excel, graficky napodobuje ztvárnění v programu *ANNIS3*. Zvýraznění buněk provedl pro přehlednost autor této práce.

Anotovány jsou pouze roviny, které jsou pro srovnání relevantní, především tedy cílové hypotézy a jejich chybové anotace.

Text je kvůli své délce segmentován na jednotlivé číslované věty (souvětí).

(1)

tok	Můj	nejhorší	den	v	životě
TH1	Můj	nejhorší	den	v	životě
TH1Diff					
TH2	Můj	nejhorší	den	v	životě
TH2Diff					
TXTStructure	start				
macro	title				

(2)

tok		Studovala	jsem		v	střední	škole	,	měla	jsem		nejhorší	den	.
TH1	Když		jsem	studovala	na	střední	škole	,	měla	jsem		nejhorší	den	.
TH1Diff	INS	DEL		INS	CHA									
TH2	Když		jsem	studovala	na	střední	škole	,	zažila	jsem	svůj	nejhorší	den	.
TH2Diff	INS	DEL		INS	CHA				CHA		INS			

(3)

tok	Ráno	jsem	vstavala	pozdě	,	protože	můj	budík	nefungoval	.
TH1	Ráno	jsem	vstávala	pozdě	,	protože	můj	budík	nefungoval	.
TH1Diff			CHA							
TH2	Ráno	jsem	vstala	pozdě	,	protože	můj	budík	nefungoval	.
TH2Diff			CHA							

(4)

tok	Stihnula	jsem		do	školy	autobusem	
TH1	Stihnula	jsem	autobus	do	školy		,
TH1Diff			DEL			INS	INS
TH2	Stihla	jsem	autobus	do	školy		,
TH2Diff	CHA		DEL			INS	INS

(5)

tok			zase	měla	jsem		dopravní	zacpu	,
TH1	ale		zase			byla	dopravní	zácpa	,
TH1Diff	INS			DEL	DEL	INS		CHA	
TH2	ale	protože	zase			byla	dopravní	zácpa	,
TH2Diff	INS	INS		DEL	DEL	INS		CHA	

(6)

tok	čekala	jsem	kolem	1	hod	.
TH1	čekala	jsem	kolem	1	hod	.
TH1Diff						
TH2	čekala	jsem	asi	1	hodinu	.
TH2Diff			CHA		CHA	

(7)

tok	Když	jsem	jela	do	školy		měli	jsme	výučovaní	už	končil	.
TH1	Když	jsem	jela	do	školy	,			vyučování	už	končilo	
TH1Diff						INS	DEL	DEL	CHA		CHA	.
TH2	Když	jsem	přijela	do	školy	,			vyučování	už	končilo	.
TH2Diff			CHA			INS	DEL	DEL	CHA		CHA	

(8)

tok	byla	jsem		špatná	.	Ale	jsem	neměla	dobrý	napad	,
TH1	Byla	jsem		špatná	.	Ani	jsem	neměla	dobrý	nápad	
TH1Diff	CHA					CHA				CHA	DEL
TH2	Bylo		to	špatné	.	Ani	jsem	neměla	dobrý	nápad	
TH2Diff	CHA	DEL	INS	CHA		CHA				CHA	DEL

(9)

tok	a	jenom	jsem	cvičila		hudební	stroje	,
TH1	a	jenom	jsem	cvičila	na	hudební	nástroje	,
TH1Diff				CHA	CHA		CHA	
TH2	a	jenom	jsem	cvičila	na	hudební	nástroje	,
TH2Diff				CHA	INS		CHA	

(10)

tok	protože		jsem	měla		vyučov	na	hudební	stroje	.
TH1	protože		jsem	měla		vyučování	na	hudební	nástroje	.
TH1Diff						CHA			CHA	
TH2		měla	jsem		totiž	vyučování	na	hudební	nástroje	.
TH2Diff	DEL	MOVT		MOVS	INS	CHA			CHA	

6.3 FalkoSummary

Anotační formát subkorpusu FalkoSummary je blíže vázán na popis německé větné struktury. Popis se využívá při anotaci na rovině deskripce. Protože nelze v češtině stejný systém využít, rovina deskripce nebude do anotace zařazena. Ačkoli bude porušen jeden z konstitučních prvků tohoto anotačního formátu (postup korespondující s postupem chybové analýzy – identifikace, deskripce, explanace) je dle mého názoru srovnání s ostatními modely přínosné, protože stěžejní anotační prvky (určení rozsahu chyby, možnost anotovat překrývající se řetězce a vícerovinný formát založený na lingvistických kategoriích) zůstávají zachovány.

Pro potřeby označení jednotlivých rovin budeme používat tyto zkratky (vycházejí z popisu korpusu v Lüdeling et al. 2006):

ORT – pravopis (ortography); WF – slovotvorba (worf formation); AGR – shoda (agreement); GOV – syntaktická závislost (government); TEN – čas (tense); MO – mood (způsob); WO – slovosled (word order); EXP – styl (expression).

Rovinu identifikační a explanační budeme rozlišovat pomocí sufixu „I“ a „E“.

Anotace je stejně jako v případě FalkoEssay provedena v programu MS Excel.

Zvýraznění buněk provedl pro přehlednost autor této práce.

Text je kvůli své délce segmentovaný na jednotlivé číslované věty (souvětí).

(1)

word	Můj	nejhorší	den	v	životě
target					

(2)

word		Studovala	jsem	v	střední	škole	,	měla	jsem	nejhorší	den	.
target	Když	jsem studovala		na								
ORT_I		x										
ORT_D		změna										
GOV_I		x										
GOV_D				změna								
WO_I		x										
WO_D		přesun										

(3)

word	Ráno	jsem	vstavala	pozdě	,	protože	můj	budík	nefungoval	.
target			vstala							
EXP_I			x							
EXP_D			změna							

(4)

word	Stihnula	jsem	do	školy	autobusem
target	Stihla		autobus do školy,		
GOV_I	x				
GOV_D					změna
WO_I			x		
WO_D			přesun		
EXP_I	x				
EXP_D	změna				

(5)

word		zase	měla	jsem	dopravní	zacpu	,
target	ale protože		byla			zácpa	,
GOV_I			x				
GOV_D						změna	
EXP_I			x				
EXP_D			změna				

(6)

word	čekala	jsem	kolem	1	hod	.
target			asi		hodinu	.
GOV_I	x					
GOV_D					změna	
EXP_I			x			
EXP_D			změna			

(7)

word	Když	jsem	jela	do	školy		měli	jsme	výučovaní	už	končil	.
target			přijela			,	vyučování už končilo					
ORT_I	x											
ORT_D						vložení						
AGR_I									x			
AGR_D											změna	
EXP_I	x						x					
EXP_D			změna				vymazání					

(8)

word	byla	jsem	špatná	.	Ale	jsem	neměla	dobrý	napad	,
target	Bylo to		špatné		Ani				nápad	
ORT_I	x								x	
ORT_D	změna								změna	
AGR_I	x									
AGR_D			změna							
EXP_I	x				x					
EXP_D	změna				změna					

(9)

word	a	jenom	jsem	cvičila	hudební	stroje
target				cvičila na hudební nástroje		
ORT_I				x		
ORT_D				změna		
GOV_I				x		
GOV_D				změna		
EXP_I						x
EXP_D						změna

(10)

word	protože	jsem	měla	vyučov	na	hudební	stroje	.
target	měla jsem totiž			vyučování			nástroje	
WO_I	x							
WO_D	přesun							
EXP_I	x			x			x	
EXP_D	změna			změna			změna	

6.4 Komentář a doplnění

Cílová hypotéza u korpusu CZESL byla určena s ohledem na jeho základní anotační pravidlo: anotátor by měl do textu co nejméně zasahovat a měl by se snažit pouze o gramatickou správnost, nikoli stylistickou vytříbenost. Podobně je tomu i u minimální cílové hypotézy (TH1) subkorpusu FalkoEssay. Maximální cílová hypotéza FalkoEssay (TH2) na druhou stranu zpracovává i stylistické a další úpravy. Cílová hypotéza FalkoSummary taktéž formuluje opravy nad rámec gramatické správnosti, vyplývá to z charakteru anotace rozdělené do osmi rovin dle lingvistických kritérií.

Ve větě 1 je ve všech anotacích doplněna spojka „když“, která uvozuje vedlejší větu. FalkoEssay řeší opravu prostým přidáním tokenu a označením chyby jako INS, tedy vložení. Není zde naznačen vztah k ostatním tokenům.

U korpusu CZESL je oprava provedena v souladu s pravidlem postupné anotace, tedy až na třetí rovině. Vztah je zde explicitně naznačen chybovým kódem *dep* (syntaktická závislost) s odkazem na slovo, které opravu motivuje, tedy přísudek. Ideální by bylo odkazovat na celou hlavní větu, protože ta především motivuje užití spojky, kvůli omezení anotačního formátu je však tato anotace nemožná.

Oprava v rámci anotace korpusu FalkoSumary je však problémová. V tomto anotačním formátu nelze tokeny přidat, lze pouze formulovat hypotézu coby jeden token sdružující přidaný výraz se zbytkem souvisejících tokenů, a poté označit, kterých originálních pozic se oprava týká. Označení tokenu, který se předtím nevyskytoval na rovině originálu, není možné, vložení by proto bylo sdruženo pod chybový tag EXP spolu s chybou ve slovosledu, jehož označení by se tak stalo nejednoznačné. Stejný případ nastává i ve větě 5 („ale protože“). Ačkoli je tato anotace dle FalkoSummary přípustná, aplikace na češtinu a její volný slovosled je přinejmenším problematická.

Podobná situace je taktéž ve větě 10 (označeno červeně). Z důvodu lepší srozumitelnosti byla spojka „protože“ zaměněna za spojku „totiž“ (došlo tak ke změně poměru mezi větami), z provedené anotace však lze jen obtížně vyčíst vztahy mezi jednotlivými tokeny.

Jako anotaci, která pomíjí některý z aspektů chyby, uvádíme doplnění čárky

(zvýrazněno červeně) na konci věty 4. Anotace je provedena správně, nelze však označit chybu v interpunkci (rovina ORT), informace je proto ztracena.

Pozitivním prvkem formátu FalkoSummary je však možnost určit jakýkoli rozsah působnosti emendace (rovina identifikace). Jak vidíme ve větě 2, při opravě předložky „na“ mohla být jako identifikace chyby označena celá fráze. Pokud porovnáme anotaci stejné věty v korpusu CZESL, zjistíme, že odkaz vede ke slovu „škola“. Užití předložky „na“ místo „v“ však ovlivňuje také řídící sloveso, k tomu však již odkaz nevede. Je otázkou, který větný člen motivuje opravu chyby více, tvůrci korpusu CZESL se však rozhodli odkazovat pouze na jeden z nich v zájmu zachování jednoduchosti anotace.

Podobný problém, který však již není v anotačním příkladu zachycen, je anotace zmnožených syntaktických pozic, které spolu nesousedí. Uspokojivě ho neřeší ani jeden z anotačních modelů. Uvažujme následující větu:

„Krásná Sněhurka, zlá Ježibaba a vždy galantní princ *stálo u chaloupky.“

Jako chyba ve shodě (rovina AGR) by se v subkorpus FalkoSummary opravil tvar „stálo“ na „stáli“ a označil by se celý několikanásobný podmět včetně rozvíjejících větných členů jako vymezení chyby.

Podle manuálu korpusu CZESL (viz Rosen a Štindlová 2012, s. 32) se vede odkaz od přísudku (označeného chybovým tagem *Agr*) pouze na první spojovací výraz v koordinační vazbě zleva či první čárku zleva, pokud spojovací výraz neexistuje. Zbytek se neoznačuje.

Na základě příkladu lze vidět, že ani jedním ze způsobů anotace není možné plně zachytit několikanásobný podmět.

U korpusu FALKO tato situace nenastává, protože chyby se označují pouze na úrovni povrchové realizace. V tomto případě by se jednotlivé části mnohonásobného podmětu anotovaly každá zvlášť, jako izolované tvary.

Dalším důležitým aspektem chybové anotace je otázka úpravy slovosledu.

U korpus CZESL probíhá značení chyb tohoto druhu automatickou analýzou, anotátor při anotaci pouze přesouvá potřebné tokeny na správná místa a žádný tag nepřiřazuje.

Korpus FalkoEssay užívá značek MOVS („move source“) a MOVT („move target“). Tokeny si nikdy nevyměňují místa, ale přesunují se, přičemž zdrojový token zůstává

prázdný, jak vyplývá z příkladu anotace, viz kapitola 6.2, věta 10, a jak je popsáno v popisu FalkoEssay, kapitola 5.3.1.1. V rámci tohoto systému se však můžeme setkat s výjimkami, viz anotaci FalkoEssay, věta 2 a 4 (vyznačeno žlutě). V prvním případě je kromě chyby ve slovosledu přítomna navíc i pravopisná chyba, v druhém případě chyba v rekci. Přiřadit dvě chybové značky nelze, je proto zvolen způsob anotace vymazáním (DEL) a nově vložením na patřičné místo (INS).

FalkoSummary při změně slovosledu nepracuje s jednotlivými tokeny. Jejich přesun řeší přepisem chybného úseku jako celku, k němuž poté přiřazuje chybový tag.

Z exemplifikace i předchozích odstavců vyplývá výhoda anotačního formátu korpusu CZESL a FalkoSummary – možnost přiřazovat k jednomu tokenu více chybových značek. Děje se tak pomocí samostatných rovin, z nichž se každá týká jednoho typu chyby (FalkoSummary), či prostým přidáním dalšího tagu na uzel mezi slovními tvary (CZESL).

Korpus CZESL má ještě jednu výhodu: možnost anotovat dílčí změny jako sekundární chybu. Jak vidíme ve větě 5, fráze „měla jsem dopravní zácpu“ je gramaticky správná, avšak významem v daném kontextu nepřijatelná. Změna pádu ve slově „zácpu“ je důsledkem změny „měla“ na „byla“, je tudíž sekundární.

Pro úplnost výčtu je třeba ještě dodat, že subkorpus FalkoEssay umožňuje anotovat makro strukturu textu. Zavádí kvůli tomu dvě roviny: TXTStructure (označuje první a poslední token v textu) a macro (označuje specifické úseky textu jako nadpis, přímá řeč apod.)

7 Závěr

Cílem práce bylo srovnat anotační modely korpusů CZESL a FALKO s ohledem na jejich chybovou anotaci.

Teoretickým srovnáním vyšlo najevo, že ačkoli oba vybrané korpusy využívají vícerovinné distanční anotace, přistupuje k ní každý jiným způsobem. Rozdíly se týkají nejen chybových taxonomií (reflektující cílový jazyk daného korpusu), ale i přizpůsobení anotačního modelu předpokládanému účelu toho kterého korpusu.

Přestože výhody vícerovinného distančního modelu, jak je prezentují autoři korpusu FALKO, jsou neoddiskutovatelné, každý anotační systém zpracovávající všeobecnou a komplexní chybovou anotaci má svá pozitiva i negativa. Vyznačit některé z nich bylo cílem exemplifikace, což znamenalo paralelní anotaci části českého žákovského textu v prostředí anotačních modelů obou korpusů.

Principy vícerovinné anotace formulované spolu s korpusem FALKO (který byl prvním takto konstruovaným žákovským korpusem) přecházejí i do anotace korpusu CZESL, anotace korpusu FALKO je však založena na možnosti přidávat anotační roviny podle potřeby anotace, zatímco korpus CZESL užívá pouze tři anotační roviny.

Tento nejvýraznější rozdíl se promítá i v charakteru korpusu, kdy FALKO akcentuje svým otevřeným formátem maximalizaci možností chybové anotace, korpus CZESL naproti tomu upřednostňuje ne tak rozsáhlý anotační model, kde je možné lépe zachytit vztahy mezi jednotlivými tokeny, což prokazuje svou důležitost v češtině a jejím volném slovosledu.

8 Pojmy a zkratky

token – výskyt slovního tvaru v korpusu; rovněž také jednotlivá pozice v korpusu (slovní tvar i interpunkce)

(podle Pala 1996)

elicitované projevy – projevy, které jsou určitou měrou řízené jinou osobou, než mluvčím samotným (dochází zde k omezení spontaneity). Projevy mohou být buďto zcela neřízené (tj. přirozené, spontánní), nebo naopak zcela řízené (tzv. experimentálně elicitované – sdělení je jasně řízeno s důrazem na formu). Mezi těmito dvěma póly je možné určit další typy projevů, závisejících na míře elicitovanosti. Příkladem mohou být klinicky elicitované projevy, při nichž je mluvčí do určité míry řízen, nicméně je kladen důraz na obsah výpovědi, nikoli jeho formu. Patří zde například práce žáků vzniklých ve výukovém kontextu. Právě na těchto datech bývají z velké části založeny žákovské korpusy (CZESL i FALKO není výjimkou).

(podle Šebesta a Škodová 2012, s. 18–19)

anotace – obecně připojování poznámek k textu, úžeji v rámci korpusové lingvistiky přiřazování morfologické či chybové značky

emendace – samotná oprava chyby

metadata – data o datech, informace strukturovaně popisující data korpusu

vkládaná anotace (inline markup) – značky jsou umístěny přímo v textu

jednorovinná anotace – značky a text korpusu jsou na jedné společné anotační rovině

distační anotace (stand-off markup) – značky jsou umístěny jinde, než v textu samotném

vícerovinná anotace – anotace probíhá na více rovinách; na každé z anotačních rovin může probíhat odlišný typ anotace

longitudinální korpus – vývojový korpus. Zaměřuje se na sledování určitého žáka anebo skupiny žáků v průběhu jeho/jejich jazykového vývoje (podle Šebesta 2010, s. 16)

průřezový korpus – zachycuje stav jazykové kompetence žáka/žáků v jediném úseku jeho/jejich jazykového vývoje (podle Šebesta 2010, s. 16)

POS (Part-of-speech) – anglicky „slovní druhy“, v korpusové lingvistice odkazuje na morfologickou anotaci (POS tag, POS tagging)

lemmatizace – přiřazování lemma

html – HyperText Markup Language; značkovací jazyk užívaný pro tvorbu webových stránek či dokumentů

tag – značka/kód, který je přiřazován při anotaci

tagset – ucelený soubor tagů/kódů/značek pro anotaci korpusu (POS tagset, chybový tagset apod.)

9 Seznam použité literatury

ANON., 2003. *The Stuttgart-Tübingen Tagset (STTS)* [online] [vid. 29. červen 2013]. Dostupné z: http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html

ANON., 2010. Komplexe Verben. *Basic German Grammar* [online] [vid. 15. červen 2013]. Dostupné z: http://cla.unipv.it/wp-content/uploads/corsicambridge/CULP_BasicGerman/grammar/content/bgg5_1.html

ANON., 2011. *EXMARaLDA* [online]. Hamburg Centre for Language Corpora [vid. 29. červen 2013]. Dostupné z: http://www.exmaralda.org/index.html

ANON., 2012. *C-Test Der Sprachtest* [online] [vid. 15. červen 2013]. Dostupné z: http://www.c-test.de/

ANON., 2013a. *A Learner Corpus of Czech* [online] [vid. 14. duben 2013]. Dostupné z: http://utkl.ff.cuni.cz/learncorp/

ANON., 2013b. *Akces | Akviziční korpusy českého jazyka* [online] [vid. 20. duben 2013]. Dostupné z: http://akces.ff.cuni.cz/

ANON., 2013c. *CZESL-PLAIN* [online] [vid. 20. duben 2013]. Dostupné z: http://korpus.cz/czesl-plain.php

ANON., 2013d. Dostupné korpusy. *Český národní korpus* [online] [vid. 14. duben 2013]. Dostupné z: http://korpus.cz/struktura.php

ANON., 2013e. *Falko (ANNIS Corpus Search)* [online] [vid. 29. červen 2013]. Dostupné z: http://korpling.german.hu-berlin.de/falko-suche/

ANON., 2013f. *Konkordance czesl-plain* [online] [vid. 1. květen 2013]. Dostupné z: https://www.korpus.cz/corpora/run.cgi/first_form?corpname=omezeni/czesl-plain

ANON., 2013g. *Popis morfologických značek* [online] [vid. 12. květen 2013]. Dostupné z: http://korpus.cz/bonito/znacky.php

ANON., 2013h. Projekt a realizační tým. *Inovace vzdělávání v oboru čeština jako druhý jazyk* [online] [vid. 14. duben 2013]. Dostupné z: http://www.c2j.cz/projekt-a-realizacni-tym

DÍAZ-NEGRILLO, Ana a Jesús FERNÁNDEZ-DOMÍNGUEZ, 2006. Error tagging systems for learner corpora. *RESLA*. roč. 2006, č. 19, s. 83–102.

DIETTERLE, Burkhard, 2013a. *Access — Korpuslinguistik und Morphologie* [online] [vid. 2. červenec 2013]. Dostupné z: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/research/falko/access

DIETTERLE, Burkhard, 2013b. *Design — Korpuslinguistik und Morphologie* [online] [vid. 2. červenec 2013]. Dostupné z: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/research/falko/design

DIETTERLE, Burkhard, 2013c. *Falko — Korpuslinguistik und Morphologie* [online] [vid. 29. červen 2013]. Dostupné z: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/research/falko/standardseite?set_language=en&cl=en

HANA, Jirka, 2012. *feat* [online]. Dostupné z: http://ufal.mff.cuni.cz/~hana/feat.html

HRDLIČKOVÁ, Tereza, 2011. *Manuál pro přepis psaných materiálů* [online] [vid. 15. červen 2013]. Dostupné z: http://utkl.ff.cuni.cz/~rosen/public/transkripce.pdf

JAMES, Carl, 1998. *Errors in Language Learning and Use : exploring error analysis*. Harlow: Longman.

JELÍNEK, Tomáš, 2008. Nové značkování v Českém národním korpusu. *Naše řeč*. roč. 91, č. 1, s. 13–20.

JIA WEI, Chan, 2007. *Falko Summary Documentation* [online] [vid. 15. červen 2013]. Dostupné z: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/research/falko/falkodata/falkokern.en/at_download/file

LÜDELING, Anke, 2007. *Falko Georgetown Dokumentation* [online] [vid. 15. červen 2013]. Dostupné z: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/pdf/BeschreibungFalkoGeorgetown.pdf

LÜDELING, Anke, Maik WALTER, Emil KROYMANN a Peter ADOLPHS, 2006. Multi-level error annotation in learner corpora. In: *Proceedings from the Corpus Linguistics Conference 2005, Birmingham* [online]. [vid. 20. duben 2013]. Dostupné z: http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-journal/LanguageLearningandError/multilevelerror.doc

PALA, Karel, 1996. Informační technologie a korpusová lingvistika (1). *Zpravodaj ÚVT MU*. roč. VI, č. 3, 11, s. 8–11.

REZNICEK, Marc, Anke LÜDELING a Hagen HIRSCHMANN, v přípravě. Competing Target Hypotheses in the Falko Corpus. A Flexible Multi-Layer Corpus Architecture. In: Ana DÍAZ-NEGRILLO *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins.

REZNICEK, Marc, Anke LÜDELING, Cedric KRUMES, Franziska SCHWANTUSCHKE, Maik WALTER, Karin SCHMIDT, Hagen HIRSCHMANN a Torsten ADREAS, 2012. *Das Falko-Handbuch, Korpusaufbau und Annotationen, Version 2.01* [online] [vid. 18. duben 2013]. Dostupné z: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen_v2.01

ROSEN, Alexandr, 2011. *Doplnění k manuálu 3* [online] [vid. 15. červen 2013]. Dostupné z: http://utkl.ff.cuni.cz/~rosen/public/ transkripce_doplnek.pdf

ROSEN, Alexandr a Barbora ŠTINDLOVÁ, 2012. *Návod k anotaci chybového korpusu* [online] [vid. 18. duben 2013]. Dostupné z: http://utkl.ff.cuni.cz/~rosen/public/anotace.pdf

SCHILLER, Anne a Simone TEUFEL, 1995. *Guidelines für das Tagging deutscher Textcorpora mit STTS* [online]. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung. [vid. 18. duben 2013]. Dostupné z: http://www.sfs.uni-tuebingen.de/Elwis/stts/stts-guide.ps.gz

SCHMID, Helmut, 2013. *Treetagger* [online] [vid. 29. červen 2013]. Dostupné z: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

SCHMID, Helmut a Florian LAWS, 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In: Scott DONIA *22nd International Conference on Computational Linguistics* [online]. Manchester, United Kingdom: Coling, s. 777–784. Dostupné z: http://dl.acm.org/citation.cfm?id=1599081.1599179.

SIMONČIČ, Jernej, 2013. *GIMP for Windows* [online]. Dostupné z: http://gimp-win.sourceforge.net/

SKOUMALOVÁ, Hana, 2013. *Pozční tagy* [online] [vid. 12. květen 2013]. Dostupné z: http://utkl.ff.cuni.cz/~skoumal/morfo/

SUPPUS, Nina, 2013. *WHiG - What's hard in German? — Korpuslinguistik und Morphologie* [online] [vid. 4. červenec 2013]. Dostupné z: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/research/learner-difficulties/

ŠEBESTA, Karel, 2010. Korpusy češtiny a osvojování jazyka. *Studie z aplikované lingvistiky*. roč. 2010, č. 2, s. 11–33.

ŠEBESTA, Karel a Svatava ŠKODOVÁ, ed., 2012. *Čeština - cílový jazyk a korpusy* [online]. Liberec: Technická univerzita v Liberci. ISBN 978-80-7372-848-9. Dostupné z: http://utkl.ff.cuni.cz/~rosen/public/czesl_monografie.pdf

ŠTINDLOVÁ, Barbora, 2011. *Evaluace chybové anotace v žákovském korpusu češtiny* [online]. Praha [vid. 20. duben 2013]. Disertační práce. Univerzita Karlova. Filozofická fakulta. Ústav českého jazyka a teorie komunikace. Dostupné z: https://is.cuni.cz/webapps/zzp/detail/25046

ŠTINDLOVÁ, Barbora, Svatava ŠKODOVÁ, Jirka HANA a Alexandr ROSEN, 2011. Víceúrovňová anotace českého žákovského korpusu. In: *Korpusová lingvistika Praha 2011: 3 – Gramatika a značkování korpusů*. Praha: Ústav Českého národního korpusu, Nakladatelství Lidové noviny, s. 208–225.

ZELDES, Amir, 2013. *EXMARaLDA Add-In für MS Excel* [online] [vid. 29. červen 2013]. Dostupné z: http://www.exmaralda.org/exceladdin.html

10 Přílohy

Příloha 1: Chybové tagy korpusu CZESL

Převzato z (Rosen a Štindlová 2012, s. 60)

typ chyby	podtyp chyby	A/M	popis chyby	počet hran nahoru: min;max	počet hran dolů: min;max	počet odkazů: min;max	rovina
incor			nesprávný tvar (obecná kategorie pro neanotované případy oprav)				
	incorInfl	M	nesprávná flexe	1;1	1;1	0;0	R1
	incorBase	M	nesprávný kmen	1;1	1;1	0;0	R1
	incorOther	A	ostatní nesprávné tvary	1;-1	1;-1	0;0	R1
fw			cizí / nově utvořené / neidentifikovatelné slovo				
	fwFab	M	nově utvořené „české“ slovo	1;1	1;1	0;0	R1
	fwNc	M	slovo z jiného jazyka	1;1	1;1	0;0	R1
flex		M	flexe u výrazů fw	1;1	1;1	0;0	R1
wbd			chybná hranice slov				
	wbdPre	M	prefix oddělený mezerou a předložka bez mezery	1;2	1;2	0;0	R1
	wbdComp	M	neoprávněně rozdělená/spojená kompozita	2;-1	1;1	0;0	R1
	wbdOther	M	jiné chyby týkající se hranice slov	1;-1	1;-1	0;0	R1
agr		M	narušení shody	1;1	1;1	0;-1	R2
dep		M	chyba ve vyjádření syntaktické závislosti	0;1	0;1	0;-1	R2
ref		M	chyba v zájmenném odkazu	1;1	1;1	0;1	R2
vbx		M	chyba v analytickém slovesném tvaru a složeném přísudku	1;-1	1;-1	0;1	R2
	cvf	A	chyba v analytickém slovesném tvaru	1;-1	1;-1	0;1	R2
	mod	A	chyba v konstrukci s modálním nebo fázovým slovesem	1;-1	1;-1	0;1	R2
	vnp	A	chyba ve sponově-jmenném přísudku (vč. pas. a rez.)	1;-1	1;-1	0;1	R2
rflx		M	chyba v reflexivním výrazu	0;-1	0;-1	0;-1	R2
neg		M	chyba v negaci	1;-1	1;-1	0;1	R2
odd		A	nadbytečné slovo	1;1	0;0	0;0	R2
miss		A	chybějící slovo	0;0	1;1	0;0	R2
wo		A	chybný slovosled	1;-1	1;-1	0;0	R2
lex		M	chyba v lexiku a frazeologii	0;-1	0;-1	0;1	R2
use		M	chyba ve významovém užití gramatické kategorie	1;-1	1;-1	0;1	R2
sec		M	sekundární, „zavlečená“ chyba	1;-1	1;-1	0;-1	R2
styl			obecněčeský, knižní, nářeční tvar				
	stylColl	M	potenciální obecněčeský tvar	0;-1	0;-1	0;-1	R1,R2
	stylOther	M	knižní, nářeční, slangový ap. tvar/výraz	0;-1	0;-1	0;-1	R1,R2
	stylMark	M	výplňková slova	1;1	0;0	0;0	R2
disr		M	rozvrácená konstrukce	-1;-1	0;-1	0;0	R2
problem		M	problémová chyba	0;-1	0;-1	0;-1	R1,R2

Příloha 2: Kompletní anotační schéma korpusu FalkoEssay

Zpracováno podle (Reznicek et al. 2012, s. 6–8), přeloženo.

název anotační kategorie	anotační rovina	popis
leaner	TXTstructure	označuje první (start) a poslední (end) token v textu
	macro	macro struktura textu (nadpis, citace, titulek ...)
	fm	označuje výskyt cizojazyčného textu
falko	word	text segmentovaný na tokeny
	pos	morfologická anotace roviny *word*
	lemma	lemmatizace roviny *word*
ctok	ctok	manuálně opravená tokenizace roviny *word*, zároveň taktéž výchozí rovina pro formulaci cílových hypotéz a referenční rovina
	ctokpos	POS roviny *ctok*
	ctoklemma	lemmatizace roviny *ctok*
TH1	TH1	minimální cílová hypotéza
	TH1Diff	chybová anotace, rozdíl mezi *TH1* a *ctok*
	TH1S	číselné označení vět
	TH1pos	morfologická anotace cílové hypotézy *TH1*
	TH1posDiff	rozdíl mezi *TH1pos* a *ctokpos*
	TH1gpos	manuálně korigovaná morfologická anotace cílové hypotézy TH1
	TH1gposDiff	rozdíl mezi manuálně korigovanou morfologickou anotací *TH1* (*TH1gpos*) a *ctokpos*
	TH1lemma	lemmatizace cílové hypotézy *TH1*
	TH1lemmaDiff	rozdíl mezi *TH1lemma* a *ctoklemma*
TH2	TH2	maximální cílová hypotéza
	TH2Diff	chybová anotace, rozdíl mezi *TH2* a *ctok*
	TH2S	číselné označení vět
	TH2pos	morfologická anotace cílové hypotézy *TH2*
	TH2posDiff	rozdíl mezi *TH2pos* a *ctokpos*
	TH2lemma	lemmatizace cílové hypotézy *TH2*
	TH2lemmaDiff	rozdíl mezi *TH2lemma* a *ctoklemma*

TH0	TH0	minimální cílová hypotéza, ve které byly vráceny přesuny tokenů
	TH0Diff	chybová anotace, rozdíl mezi *TH0* a *ctok*
	TH0S	číselné označení vět
	TH0pos	morfologická anotace cílové hypotézy *TH0*
	TH0posDiff	rozdíl mezi *TH0pos* a *ctokpos*
	TH0gpos	manuálně korigovaná morfologická anotace cílové hypotézy TH0
	TH0gposDiff	rozdíl mezi manuálně korigovanou morfologickou anotací *TH0* (*TH0gpos*) a *ctokpos*
	TH0lemma	lemmatizace cílové hypotézy *TH0*
	TH0lemmaDiff	rozdíl mezi *TH0lemma* a *ctoklemma*
THverb	THverb	cílová hypotéza komplexních sloves
	THverbDiff	chybová anotace, rozdíl mezi *THverb* a *ctok*
	THverbS	číselné označení vět
	THverbpos	morfologická anotace cílové hypotézy *THverb*
	THverbposDiff	rozdíl mezi *THverbpos* a *ctokpos*
	THverblemma	lemmatizace cílové hypotézy *THverb*
	THverblemmaDiff	rozdíl mezi *THverblemma* a *ctoklemma*
	verbkategorie	rozlišení: prefix vs
	verbform	morfosyntaktická forma ve větě
	verblemma	lemma složeného slovesa
	verbfehlertyp	klasifikace chyby pro komplexní slovesa, každá chyba jeden tag
	verbfehlertyp_all	klasifikace chyby pro složená slovesa, všechny chyby na jednom tokenu
dep	dep	závislosti na rovině TH1
	func	gramatická funkce na TH1 podle Foth, Kilian A. (2006): *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*. Universität Hamburg.

Příloha 3: Anotace komplexních sloves

Zpracováno podle (Reznicek et al. 2012, s. 63–66), upraveno a přeloženo

anotační rovina	tag	popis
verbkategorie	vpart	Partikelverb, slovesná část
	ppart	samostatně stojící neverbální část Partikelverb
	vpräf	Präfixverb
	ppräf	samostatně stojící prefix (tedy jedná se o chybu)
	vpartx	místo, na kterém mělo stát Partikelverb, ale nestojí
	ppartx	místo, na kterém měla stát neslovesná část Partikelverb, ale nestojí
	vpräfx	místo, na které mělo stát Präfixverb, ale nestojí
verblemma	-	určení lemmatu
verbfehlertyp	sem	verblemma existuje, ale je špatně užito (správné musí být uvedeno jako cílová hypotéza)
	orth	pravopisná chyba
	lex	verblemma neexistuje (správné musí být uvedeno jako cílová hypotéza)
	part	nadužití neslovesné části Partikelverb
	as	chyba v argumentační struktuře
	flex	chyba ve flexi
	ge	chyba ve špatném užití či vynechání předpony „ge“
	zu	chyba ve špatném užití předpony „zu“
	ws	chyba ve slovosledu komplexního slovesa
	sep	špatné rozdělení komplexního slovesa
verbform	fin	finitní tvar
	finsep	finitní tvar, syntakticky oddělený
	inf	infinitiv bez „zu“
	infzu	infinitiv s „zu“
	p2	participium II
	p1	participium I
	nn	speciální případ, kdy se nejedná o sloveso

Příloha 4: Kompletní anotační schéma korpusu FalkoSummary

Zpracováno podle (Reznicek et al. 2012, s. 5–6)

název anotační kategorie	anotační rovina	popis
	word	text segmentovaný na tokeny
	pos	morfologická anotace roviny *word*
	lemma	lemmatizace roviny *word*
	target hypothesis	cílová hypotéza
	cpos	manuální POS tagy
	transcriptor comment	komentář přepisovače
Topologická pole	matrix-satz	Martrixsatz 1
	matrix-satz_felder	Martrixsatz 1: topologische Felder
	konstituenten-satz_1	Konstituentensatz 1
	konstituenten-satz_1_ felder	KS1: topologische Felder
	konstituenten-satz_1_ felder_2	KS1: topologische Felder
	matrix-satz_2	Martrixsatz 2
	konstituenten-satz_2	Konstituentensatz 2
	konstituenten-satz_2_felder	KS2: topologische Felder
	konstituenten-satz_2_ felder_2	KS2: topologische Felder
	konstituenten-satz_3	Konstituentensatz 3
	konstituenten-satz_3_ felder	KS3: topologische Felder
	konstituenten-satz_3_ felder_2	KS3: topologische Felder
Syntaktický popis	syntax_description_1	Syntaktische Beschreibung 1
	syntax_classification_1	Syntaktische Klassifikation 1
	syntax_classification_pos_1	Syntaktische Klassifikation POS 1
	syntax_hypothesis_1	syntaktische Zielhypothese
	syntax_description_2	Syntaktische Beschreibung 2
	syntax_classification_2	Syntaktische Klassifikation 2
	syntax_classification_pos_2	Syntaktische Klassifikation POS 2

Příloha 5: Seznam tagů STTS užívaných při automatické POS anotaci

Převzato z (Schiller a Teufel 1995)

POS =	Beschreibung	Beispiele
ADJA	attributives Adjektiv	*[das] große [Haus]*
ADJD	adverbiales oder prädikatives Adjektiv	*[er fährt] schnell* *[er ist] schnell*
ADV	Adverb	*schon, bald, doch*
APPR	Präposition; Zirkumposition links	*in [der Stadt], ohne [mich]*
APPRART	Präposition mit Artikel	*im [Haus], zur [Sache]*
APPO	Postposition	*[ihm] zufolge, [der Sache] wegen*
APZR	Zirkumposition rechts	*[von jetzt] an*
ART	bestimmter oder unbestimmter Artikel	*der, die, das,* *ein, eine*
CARD	Kardinalzahl	*zwei [Männer], [im Jahre] 1994*
FM	Fremdsprachliches Material	*[Er hat das mit "]* *A big fish [" übersetzt]*
ITJ	Interjektion	*mhm, ach, tja*
KOUI	unterordnende Konjunktion mit "zu" und Infinitiv	*um [zu leben],* *anstatt [zu fragen]*
KOUS	unterordnende Konjunktion mit Satz	*weil, daß, damit,* *wenn, ob*
KON	nebenordnende Konjunktion	*und, oder, aber*
KOKOM	Vergleichspartikel, ohne Satz	*als, wie*
NN	normales Nomen	*Tisch, Herr, [das] Reisen*
NE	Eigennamen	*Hans, Hamburg, HSV*
PDS	substituierendes Demonstrativ–pronomen	*dieser, jener*
PDAT	attribuierendes Demonstrativ–pronomen	*jener [Mensch]*
PIS	substituierendes Indefinit–pronomen	*keiner, viele, man, niemand*
PIAT	attribuierendes Indefinit–pronomen ohne Determiner	*kein [Mensch],* *irgendein [Glas]*
PIDAT	attribuierendes Indefinit–pronomen mit Determiner	*[ein] wenig [Wasser],* *[die] beiden [Brüder]*
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, dir*
PPOSS	substituierendes Possessiv–pronomen	*meins, deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein [Buch], deine [Mutter]*
	Relativpronomen	
PRELS	substituierend	*[der Hund,] der*
PRELAT	attribuierend	*[der Mann ,] dessen [Hund]*

POS =	Beschreibung	Beispiele
	Relativpronomen	
PRF	reflexives Personalpronomen	*sich, einander, dich, mir*
PWS	substituierendes Interrogativpronomen	*wer, was*
PWAT	attribuierendes Interrogativpronomen	*welche [Farbe], wessen [Hut]*
PWAV	adverbiales Interrogativ– oder Relativpronomen	*warum, wo, wann, worüber, wobei*
PAV	Pronominaladverb	*dafür, dabei, deswegen, trotzdem*
PTKZU	"zu" vor Infinitiv	*zu [gehen]*
PTKNEG	Negationspartikel	*nicht*
PTKVZ	abgetrennter Verbzusatz	*[er kommt] an, [er fährt] rad*
PTKANT	Antwortpartikel	*ja, nein, danke, bitte*
PTKA	Partikel bei Adjektiv oder Adverb	*am [schönsten], zu [schnell]*
TRUNC	Kompositions–Erstglied	*An– [und Abreise]*
VVFIN	finites Verb, voll	*[du] gehst, [wir] kommen [an]*
VVIMP	Imperativ, voll	*komm [!]*
VVINF	Infinitiv, voll	*gehen, ankommen*
VVIZU	Infinitiv mit "zu", voll	*anzukommen, loszulassen*
VVPP	Partizip Perfekt, voll	*gegangen, angekommen*
VAFIN	finites Verb, aux	*[du] bist, [wir] werden*
VAIMP	Imperativ, aux	*sei [ruhig !]*
VAINF	Infinitiv, aux	*werden, sein*
VAPP	Partizip Perfekt, aux	*gewesen*
VMFIN	finites Verb, modal	*dürfen*
VMINF	Infinitiv, modal	*wollen*
VMPP	Partizip Perfekt, modal	*[er hat] gekonnt*
XY	Nichtwort, Sonderzeichen enthaltend	*D2XW3*
$,	Komma	,
$.	Satzbeendende Interpunktion	. ? ! ; :
$(	sonstige Satzzeichen; satzintern	– [ ]()